IN PARTNERSHIP WITH:
**CNRS**

**Université Rennes 1**

Activity Report 2019

# Project-Team DYLISS

## Dynamics, Logics and Inference for biological Systems and Sequences

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

RESEARCH CENTER
**Rennes - Bretagne-Atlantique**

THEME
**Computational Biology**

# Table of contents

# Project-Team DYLISS

*Creation of the Team: 2012 January 01, updated into Project-Team: 2013 July 01*

**Keywords:**

**Computer Science and Digital Science:**

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, Visitors, External Collaborators

**Research Scientists**

Samuel Blanquart [Inria, Researcher]
François Coste [Inria, Researcher]
Anne Siegel [CNRS, Senior Researcher, HDR]

**Faculty Members**

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

**Post-Doctoral Fellows**

Celia Biane-Fourati [Inria, Post-Doctoral Fellow]
Clémence Frioux [CNRS, Post-Doctoral Fellow, until May 2019]
Vijay Ingalalli [Univ de Rennes I, Post-Doctoral Fellow, until Jan 2019]

**PhD Students**

Arnaud Belcour [Inria, PhD Student, from Oct 2019]
Lucas Bourneuf [Univ de Rennes I, PhD Student, until Nov 2019]
Mael Conan [Univ de Rennes I, PhD Student, UMR IRSET Lab]
Olivier Dennler [INSERM, PhD Student, from Oct 2019]
Nicolas Guillaudeux [Univ de Rennes I, PhD Student]
Virgilio Kmetzsch Rosa E Silva [Inria, PhD Student, from Oct 2019]
Marine Louarn [Univ de Rennes I, PhD Student]
Hugo Talibart [Univ de Rennes I, PhD Student]
Pierre Vignet [Univ de Rennes I, PhD Student, UMR IRSET Lab]
Meline Wéry [Univ de Rennes I, PhD Student, SANOFI R&D]

**Technical staff**

Méziane Aite [CNRS, Engineer]
Arnaud Belcour [CNRS, Engineer, until Jul 2019]
Xavier Garnier [Inria, Engineer]
Jeanne Got [CNRS, Engineer]

**Administrative Assistant**

Marie Le Roic [Inria, Administrative Assistant]

**Visiting Scientist**

Oumarou Abdou Arbi [Université de Maradi, Niger, until Feb 2019]

**External Collaborators**

François Moreews [INRA]
Denis Tagu [INRA, HDR]
Nathalie Théret [INSERM, HDR]

# 2. Overall Objectives

## 2.1. 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, 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 Sec. 3.2) (2) Understanding metabolism (see Sec. 3.3) (3) Characterizing regulatory and signaling phenotypes (see Sec. 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 [more info].

# 3. Research Program

## 3.1. Computer science – 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) [58], 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 [67]. 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** [57]. 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 [44]. 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* [68], 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 [url]. *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 [package]. (2) The PowerGrasp python package implements graph compression methods oriented toward visualization, and based on power graph analysis [package]. (3) The iggy package enables the repairing of an interaction graph with respect to expression data. [Python package]

## 3.3. Metabolism: from enzyme 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 enzymatic 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 [50] 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 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 [56]. 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 [65].

### *3.3.2. Associated software tools*

**Protomata**[url] 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 [50] to automate its updade and was used for refining the classification of HAD enzymes [6].

**AuReMe workspace** is designed for tractable reconstruction of metabolic networks [url]. 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: [url]. It also generated outputs to explore resulting networks with Askomics. It has been used for reconstructing metabolic networks of micro and macro-algae [62], extremophile bacteria [52] and communities of organisms [4].

**Mpwt** is a Python package for running Pathway Tools [url] on multiple genomes using multiprocessing. Pathway Tools is a comprehensive systems biology software system that is associated with the BioCyc database collection [url]. 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.

### *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) [46]. 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 [66].

**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 [47], and multi-layer information in large-scale Petri nets [43]. 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 [url]. 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) [url]. 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 [url] It is based on Guarded transition semantic and allows synchronization events to be investigated in large-scale biological networks [43]. 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

## 4.1. Application fields in biology

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.

**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 [62], [53]. 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* [55] 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 $\beta$-alanine and vitamin B5 to the seaweed via the phosphopantothenate biosynthesis pathway [63].

**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 [48] and in Salmon pathogenicity [52].

**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*) [47] and on semantic-based analysis of metabolic networks [45] was very valuable for interpreting differences of gene expression in pork meat [60] and figure out the main gene-regulators of the response of porks to several diets [59]. 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-$\beta$ signalling network.** TGF-$\beta$ is a multifunctional cytokine that regulates mammalian development, differentiation, and homeostasis with both benefical anti-tumor effect [49] and pro-tumor effect [61]. 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-$\beta$ and interleukine signaling pathways [43], [51].

# 5. Highlights of the Year

## 5.1. Highlights of the Year

The AuReMe software for metabolic network reconstruction has been selected for the Service Delivery Plan of the French Institute of Bioinformatics (IFB).

### *5.1.1. Awards*

Lucas Bourneuf is the World champion of man vs. machine challenge of the Angry Birds AI competition (during IJCAI) where humans can challenge the four best AI agents. Note that Lucas was the human, not the AI and that there is no direct connection with his PhD project. World champion nonetheless!

Nicolas Guillaudeux (with Grégoire Siekaniec from the GenScale team) won the public's prize at the short scientific film festival "Sciences en cour[t]s" for their movie about Nicolas's PhD thesis.

# 6. New Software and Platforms

## 6.1. AskOmics

*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: (1) migration to github, (2) complete re-engineering for cleaning the successive layers and accomodating further extensions, (3) integration of the Corese triplestore (https://corese.inria.fr/) in addition to fuseki and virtuoso, (4) improved user interface, (5) capability to save queries for sharing and reusing them, (6) automatic generation of askomics-compliant graph of entity types (abstraction), (7) capability to use askomics to query remote endpoints (including Uniprot and neXtProt), (8) support for federated queries involving remote endpoints and local data

- Authors: Charles Bettembourg, Xavier Garnier, Anthony Bretaudeau, Fabrice Legeai, Olivier Dameron, Olivier Filangi and Yvanne Chaussin
- Partners: Université de Rennes 1 - CNRS - INRA
- Contact: Olivier Dameron
- URL: https://github.com/askomics/askomics

## 6.2. AuReMe

*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 FUNCTIONAL DESCRIPTION: - 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

- Participants: Marie Chevallier, Meziane Aite, Guillaume Collet, Nicolas Loira, Sylvain Prigent, Jeanne Cambefort, Anne Siegel and Alejandro Maass
- Partner: University of Chile
- Contact: Meziane Aite
- Publication: Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models
- URL: http://aureme.genouest.org/

## 6.3. biseau

KEYWORDS: ASP - Answer Set Programming - Graph - Formal concept analysis

SCIENTIFIC DESCRIPTION: Use ASP as a Domain Specific Language to specify dot-based visualizations.

NEWS OF THE YEAR: First release.

- Contact: Lucas Bourneuf
- Publication: An Answer Set Programming Environment for High-Level Specification and Visualization of FCA
- URL: https://gitlab.inria.fr/lbourneu/biseau

## 6.4. 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.

- Contact: Anne Siegel
- Publication: Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species
- URL: https://github.com/AuReMe/metage2metabo

## 6.5. 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 difficulties 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 FUNCTIONAL DESCRIPTION: 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.

- Participants: Arnaud Belcour, Jacques Nicolas, Gabriel Markov and Anne Siegel
- Partner: Station Biologique de Roscoff
- Contact: Anne Siegel
- Publication: Inferring biochemical reactions and metabolite structures to cope with metabolic pathway drift
- URL: https://github.com/pathmodel

## 6.6. CADBIOM

*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)

- Participants: Geoffroy Andrieux, Michel Le Borgne, Nathalie Theret, Nolwenn Le Meur, Pierre Vignet and Anne Siegel
- Contact: Anne Siegel
- URL: http://cadbiom.genouest.org

# 7. New Results

## 7.1. Scalable methods to query data heterogenity

**Participants:** Emmanuelle Becker, Lucas Bourneuf, Olivier Dameron, Xavier Garnier, Vijay Ingalalli, Marine Louarn, Yann Rivault, Anne Siegel.

**Increasing life science resources re-usability using Semantic Web technologies** [*E. Becker, O. Dameron, X. Garnier, V. Ingallali, M. Louarn, Y. Rivault, A. Siegel* ] [25], [18], [29], [31], [23], [27], [28]. 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.

- We followed on 2018 action exploratoire Inria by studying possible optimizations for federated SPARQL queries [31]
- We considered a case-study in systems biology ([Regulatorycircuits link]), which provides tissue-specific regulatory interaction networks to elucidate perturbations across complex diseases. We relied on this structure and used Semantic Web technologies (i) to integrate the Regulatory Circuits data, and (ii) to formalize the analysis pipeline as SPARQL queries. Our result was a 335,429,988 triples dataset on which two SPARQL queries were sufficient to extract each single tissue-specific regulatory network.

- A second case-study concerned public health data for reusing electronic health data, selecting patients, identifying specific events and interpreting results typically requires biomedical knowledge [64]. We developed the queryMed R package [18], [29]. It aims to facilitate the integration of medical and pharmacological knowledge stored in formats compliant with the Linked Data paradigm (e.g. OWL ontologies and RDF datasets) into the R statistical programming environment. We showed that linking a medical database of 1003 critical limb ischemia (CLI) patients to ontologies allowed us to identify all the drugs prescribed for CLI and also to detect one contraindicated prescription for one patient. We also investigated temporal models of care sequences for the exploration of medico-administrative data as part of Johanne Bakalara's PhD, supervized with Thomas Guyet (Lacodam) and Emmanuel Oger (Repères).
- We pursued the development of AskOmics [27]. Version 3 adds the capability to generate the graph of entity types (aka abstraction) from typed RDF datasets, improved management of entity hierarchies and support for federated queries on external SPARQL endpoints such as UniProt and neXtProt.

**Graph compression and analysis** [*L. Bourneuf*] [26], [24]. Because of the increasing size and complexity of available graph structures in experimental sciences like molecular biology, techniques of graph visualization tend to reach their limit.

- We developed the Biseau approach, a programming environment aiming at simplifying the visualization task. Biseau takes advantage of Answer Set Programming and shows as a use-case how Formal Concept Analysis can be efficiently described at the level of its properties, without needing a costly development process. It reproduces the core results of existing tools like LatViz or In-Close.
- We formalized a graph compression search space in order to provide approximate solutions to the NP-complete problem of computing a lossless compression of the graph based on the search of cliques and bicliques. Our conclusion is that the search for graph compression can be usefully associated with the search for patterns in a concept lattice and that, conversely, confusing sets of objects and attributes brings new interesting problems for FCA.

## 7.2. Metabolism: from enzyme sequences to systems ecology

**Participants:** Méziane Aite, Arnaud Belcour, Mael Conan, François Coste, Clémence Frioux, Jeanne Got, Anne Siegel, Hugo Talibart.

**Modelling proteins with long distance dependencies** [*F. Coste, H. Talibart*] [15], [30], [30]

- We proposed to use information on protein contacts to train probabilistic context-free grammars representing families of protein sequences. We developed the theory behind the introduction of contact constraints in maximum-likelihood and contrastive estimation schemes and implemented it in a machine learning framework for protein grammars. Evaluation showed high fidelity of grammatical descriptors to protein structures, improved precision in recognizing sequences and the ability to model a meta-family of proteins that could not be modeled by classical approaches [15].
- We then investigated the problem of modeling proteins with crossing dependencies. Motivated by their success on contact prediction, we propose to use Potts models for the purposes of modeling proteins and searching. We developed ComPotts a tool for optimal alignment and comparison of Potts models, enabling to take into account the coevolution of residues for the search of protein homologs [30], [40].

**Large-scale eukaryotic metabolic network reconstruction** [*A. Siegel, C. Frioux, M. Aite, A. Belcour, J. Got, N. Théret, M. Conan*] [17], [14], [38]. Metabolic network reconstruction has attained high standards but is still challenging for complex organisms such as eukaryotes.

- *Large-scale eukaryotic metabolic network reconstruction*: We participated to the reconstruction of a genome–scale metabolic network for the brown Algae *Saccharina japonica* and *Cladosiphon okamuranus* in order to shed light of the specificities on the carotenoid biosynthesis Pathway.

- *Metabolic pathway inference from non genomic data*: We designed methods for the identification of metabolic pathways for which enzyme information is not precise enough. As an application study, we focused on Heterocyclic Aromatic Amines (HAAs), which are environmental and food contaminants classified as probable carcinogens. Our approach based on a refinement of molecular predictions with enzyme activity scores allows to accurately predict HAAs biotransformation and their potentials DNA reactive compounds [54].

**Systems ecology: design of microbial consortia** [*C. Frioux, A. Belcour, J. Got, M. Aite, A. Siegel*] [21], [22], [34], [33].

- We participated to the application of our methods to algal-microbial consortia, with good preliminary results, and presented them as an invited conference [22].

## 7.3. Regulation and signaling: detecting complex and discriminant signatures of phenotypes

**Participants:** Catherine Belleannée, Célia Biane-Fourati, Samuel Blanquart, Olivier Dameron, Maxime Folschette, Nicolas Guillaudeux, Marine Louarn, François Moreews, Anne Siegel, Nathalie Théret, Pierre Vignet, Méline Wéry.

**Creation of predictive functional signaling networks** [*M. Folschette, N. Théret*] [16].

- 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.

**Experimental evidences of transcript predictions** [*C. Belleannée, S. Blanquart, N. Guillaudeux*] [13].

- We designed comparative-genomics based models of gene structures through genes comparisons across species. These models enable to predict putative transcript isoforms in a species given the knowledge available in other species [46]. We recently published a first experimental validation of such a predicted transcriptome [13]. In this work, transcript isoforms of the human TRPM8 gene yield transcript predictions in the mouse TRPM8 gene, which are experimentally validated using targeted PCR in mouse tissues. This work also provides a first attempt to estimate origin of new isoforms during the gene evolution.
- In another collaboration with IGDR, we considered a multi-species gene comparison including human, mouse and dog [39]. This work reveals global trends of the gene isoform sets evolution, suggesting a extremely high plasticity of alternative transcription and alternative splicing propensities in those three species. This work moreover provides experimental evidences of the predicted transcripts based on public RNAseq data, highlighting the tissue specificity of isoform expression across species.

**Formalizing and enriching phenotype signatures using Boolean networks** [*C. Biane-Fourati, M. Wéry, A. Siegel, O. Dameron*] [20], [42], [22]

- We used Formal Concept Analysis as a symbolic bi-clustering techniques to classify and sort the steady states of a Boolean network according to biological signatures based on the hierarchy of the roles the network components play in the phenotypes. We applied our approach on a T helper lymphocyte (Th) differentiation network with a set of signatures corresponding to the sub-types of Th. This led to the identification and prediction of a new hybrid sub-type later confirmed by the literature.

# 8. Bilateral Contracts and Grants with Industry

## 8.1. Bilateral Contracts with Industry

### *8.1.1. SANOFI: co-supervised PhD*

**Participant:** Emmanuelle Becker.

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 [42]. **CIFRE co-supervised Grant: Ph.D. funding. 2017-2020**

### *8.1.2. Theranexus: co-supervised internship*

**Participant:** Pierre Beaudier.

This collaboration project was focused on assessing public databases' relevance for predicting potential drug combinations in central nervous system's pathologies [32]. It opened the perspective of a CIFRE PhD with Insiliance (under review by ANRT) **Theranexus funding. 2019**

# 9. Partnerships and Cooperations

## 9.1. Regional Initiatives

### *9.1.1. MoDaL (Brittany and Pays de la Loire regions)*

**Participant:** Olivier Dameron.

The MoDaL project is a federated project funded by BioGenOuest (Région Bretagne-Pays de la Loire) project involving scientists and engineers from IRISA/inria rennes (Genouest, Empenn, Dyliss) and the Institut du Thorax lab in Nantes. The project aims to decompartmentalize the resources dedicated to biomedical imaging and genetics. MoDaL focuses on i) establishing an inventory of the actors and infrastructures available at the inter-regional level, ii) proposing technological demonstrators that address the management, analysis and reuse of multi-infrastructure data (in-vivo, in-vitro and genomic imaging).

2019-2020. Total grant (hosted in the Empenn team): 100,000€.

### *9.1.2. PhenoMiR (European Maritime and Fisheries Fund)*

**Participant:** Emmanuelle Becker.

The PhenoMiR project is a collaboration between Fishes Physiology and Genomics Laboratory (LPGP - INRAE), eight other laboratories of the INRAE, and the Dyliss team. Its objective is first (i) to settle the first complete repository of microRNAs for the trout, exploring different physiological and breeding conditions, and then (ii) to study the potentiality of some micro-RNAs to act as bio-markers of trout breeding and development condition. The Dyliss team is responsible for the design and development of the analysis pipeline of the genomics data, including the search of potential bio-markers.

2019-2022. Total grant: 495,000€. Dyliss grant: 33,000€.

### *9.1.3. UBIQUITIN*

**Participant:** Emmanuelle Becker.

The Ubiquitin project is a collaboration between G. Rabut's team at the Institute of Developmental Biology of Rennes and the Dyliss team. It was funded as a cross-disciplinary emerging project by the University of Rennes 1.

G. Rabut's team is developing a new method to detect weak affinity protein-protein interactions based on protein complemantation with Luciferase. However the method may generate a very noisy signal depending on the *in-vivo* concentration of the partners. In the Ubiquitin project, we developped a R workflow to separate the signal from the noise in the experiments. As an application, this allowed us to decipher the intricate interplay between E2 and E3 enzymes during ubiquitination process in Yeast. This work was done during a master 2 internship. We are now continuing the project in two directions : (i) comparing the interactions identified with previously known databases, using web semantic technologies and ontologies describing protein interaction detection methods, and (ii) using formal classification to understand the structural properties of E2 and E3 that lead to their interactions. 2019-2020. Total grant: 7,000€. Dyliss grant: 4,200€.

### *9.1.4. Ph.D. fundings from Université, Inria Rennes and Inserm*

The team benefits from PhD. theses fundings by Univ. Rennes (L. Bourneuf, 2016-2019 – H. Talibart, 2017-2020 – N. Guillaudeux, 2018-2021), by Inria (A. Belcour, 2019-2022 – V. Kmetzsch, 2019–2022), by Inserm (M. Louarn, 2017-2020, Inria-Inserm PhD Grant program), and by our collaborators from IRSET (M. Conan, 2017-2020 – P. Vignet, 2018-2020, O. Dennler, 2019–2022).

## 9.2. National Initiatives

### *9.2.1. IDEALG (ANR/PIA-Biotechnology and Bioresource)*

**Participant:** Méziane Aite.

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) [More details]. 2012–20. Total grant: 11M€. Dyliss grant: 534k€.

### *9.2.2. TGFSysBio (ITMO Cancer)*

**Participant:** Olivier Dameron.

Partners are INSERM (coordinator) (IRSET, Univ. Rennes 1) CNRS (Dyliss team) and Inria (Antique, Paris). The TGFSYSBIO project aims at developing the first model of extracellular and intracellular TGF-beta system by combining a ruled-based modelling approach (kappa) and a Petri net modelling approach (cadbiom). 2015–18, extended in 2019. Total grant: 418k€. Dyliss grant: 129k€.

### *9.2.3. Programs funded by Inria*

#### *9.2.3.1. IPL Neuromarkers*

**Participant:** Emmanuelle Becker.

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.

#### *9.2.3.2. Askomics (ADT)*

**Participant:** Olivier Dameron.

AskOmics [url] 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.

## 9.3. European Initiatives

### *9.3.1. Collaborations in European Programs, Except FP7 & H2020*

- Program: Polish National Science Center
- Project acronym: NCN 2016/21/B/ST6/02158
- Project title: Grammatical inference methods in classification of amyloidogenic proteins
- Duration: January 2017 - January 2020
- Coordinator: Olgierd Unold, Politechnika Wroclawska
- Other partners: Politechnika Wroclawska (Polland)
- Abstract: The objective is to develop the methods for induction of context-free and probabilistic grammars to describe a language matching amyloidogenic protein sequences.

### *9.3.2. Collaborations with Major European Organizations*

Partner: Potsdam (Germany)

Title: Modeling combinatorial and hybrid optmization problems with Answer Set Programming

## 9.4. International Initiatives

### *9.4.1. Informal International Partners*

We have a cooperation with Univ. of Chile (MATHomics, A. Maass) on methods for the identification of biomarkers and software for biochip design. It aims at combining automatic reasoning on biological sequences and networks with probabilistic approaches to manage, explore and integrate large sets of heterogeneous omics data into networks of interactions allowing to produce biomarkers, with a main application to biomining bacteria. The program is co-funded by Inria and CORFO-chile from 2012 to 2016. In this context, Integrative-BioChile was an Associate Team between Dyliss and the Laboratory of Bioinformatics and Mathematics of the Genome hosted at Univ. of Chile funded from 2011 to 2016. The collaboration is now supported by Chilean programs.

## 9.5. International Research Visitors

### *9.5.1. Visits of International Scientists*

- **Niger:** Oumarou Abdou-Arbi (University of Maradi)

*9.5.1.1. Research Stays Abroad*

- **Germany:** Maël Conan visited the Zentrum für Bioinformatik at Hamburg University with Prof. Johannes Kirchmair for 2 months. During this stay, he learned how to predict metabolism sites using the tool developped in Pr. Kirchmair unit (FAME2/FAME3) and initiated the development of a new method to predict xenobiotics metabolism.
- **Germany:** Clémence Frioux visited the lab of Oliver Ebenoh (Heidelberg) for one week.

# 10. Dissemination

## 10.1. Promoting Scientific Activities

### *10.1.1. Scientific Events: Selection*

*10.1.1.1. Member of the Conference Program Committees*

- 12th international conference Semantic Web and Tools for Healthcare and Life Sciences (SWAT4HCLS) [O. Dameron]

- 19th NETTAB-BBCC joint international conference [E. Becker, O. Dameron]
- CSBio2019: The 10th International Conference on Computational Systems-Biology and Bioinformatics [A. Siegel]
- WBC@ICML 2019, ICML 2019 Workshop on Computational Biology [A. Siegel]
- CP'2019 25th International Conference on Principles and Practice of Constraint Programming [A. Siegel]
- Journée IA et santé, organisée par l'AFIA (Assoc. Fr. pour l'Intelligence Artificielle) et l'AIM (Assoc. Fr. d'Informatique Médicale) [O. Dameron]
- JOBIM 2019 [A. Siegel]

*10.1.1.2. Reviewer*

- Semantic Web and Tools for Healthcare and Life Sciences: O. Dameron
- Journée IA et santé: O. Dameron

*10.1.1.3. Jury member*

- Jury of interpretability prize of Sigmorphon shared task 2019: F. Coste

### *10.1.2. Journal*

*10.1.2.1. Member of the Editorial Boards*

- Journal of Biomedical Semantics [O. Dameron]
- Editor of special issue in grammatical inference of Machine Learning: [F. Coste]

*10.1.2.2. Reviewer - Reviewing Activities*

- Bioinformatics [A. Siegel, E. Becker]
- Biosystems [A. Siegel]
- Briefings in Bioinformatics [O. Dameron]
- IEEE/ACM TCBB [A. Siegel]
- Neuroinformatics [O. Dameron]
- Machine Learning [F. Coste]
- PLoS Computational biology [A. Siegel]

*10.1.2.3. Peer Community in Genomics*

In addition to the traditional publishing activities, Dyliss has also been active in the alternative approach "Peer Community in" (PCI). It is is a non-profit community of researchers who review and recommend for free (we also do it for free with traditional journals and conferences anyway), unpublished preprints in their field (i.e. unpublished articles deposited on open online archives like arXiv.org and bioRxiv.org). To a lesser extent, they may also recommend articles already published in journals.

Denis Tagu has been one of the creators of Peer Community in Genomics (PCI Genomics [More details]), which has been launched in October 2019. The scope of this PCI encompasses all aspects of genomics (structural genomics, functional genomics, epigenomics, evolutionary genomics, population genomics, proteomics, bioinformatics) dealing with every type of organisms (viruses, bacteria, fungi, plants, animals,...) as well as metagenomes.

- Denis Tagu is a manager
- Olivier Dameron is a recommender

### *10.1.3. Invited Talks*

- Dusseldorf, Quantitative Theoretical Biology group, Heinrich Heine Universitat, february 2019 [C. Frioux]
- Colloquium J. Morgenstern (Nice), february 2019 [A. Siegel]
- IGNITE International Training Network Spring school [O. Dameron]
- Colloque d'ouverture 50 ans du Laboratoire Jacques-Louis Lions, Roscoff, march 2019 [A. Siegel]
- CNRS GdR MaDics, june 2019 [O. Dameron]
- Boolean Weekend, Burlington, US, april 2019 [A. Siegel]
- Department of Plant Science, University of Oxford, may 2019 [C. Frioux]
- Quadram Institute, may 2019 [C. Frioux]
- Network analysis to elucidate natural system dynamics, diversity and performance, CECAM Conference, Lyon, may 2019 [A. Siegel]
- Séminaire IMBAS Inra Toulouse, june 2019 [A. Siegel]
- CNRS GdR BIM, november 2019 [O. Dameron]
- The 10th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2019) [N. Théret] Nice (France) December 4 to 7, 2019

### *10.1.4. Leadership within the Scientific Community*

- Member of the steering committee of the International Conference on Grammatical Inference [F. Coste]

### *10.1.5. Scientific Expertise*

*10.1.5.1. International expertise*

- Luxembourg, FNR. Core program [A. Siegel]

*10.1.5.2. Prospective working groups*

- "Big & Open Data en recherche à l'horizon 2040" foresight working group of PROSPER network [F. Coste]

*10.1.5.3. National responsibilities*

- **Institutional boards for the recruitment and evaluation of researchers**.
  - National Council of Universities, section 65 [O. Dameron, nominated member].
- **Recruitment committees**.
  - Inria Senior Researchers [national committee, A. Siegel]
  - Assistant professor in Mathematics for Biology, section 26, Univ Rennes 1 [E. Becker]

*10.1.5.4. National scientific boards*

- Animation of the Systems Biology working group of national infrastructure GDR IM and GDR BIM [A. Siegel].
- Board of directors of the French Society for biology of the extracellular matrix [N. Théret].

*10.1.5.5. Local responsibilities*

- Scientific Advisory Board of Biogenouest [N. Théret]
- Delegate to research integrity at the University of Rennes 1 [N. Théret]
- Head of the "Data and Knowledge Managment" Department (6 teams) of the IRISA lab [A. Siegel, since june 2019]
- IRISA laboratory (computer science department of Univ. Rennes 1) council [A. Siegel, until june 2019]

- Member of the Inria Rennes center council [J. Got]
- Responsability of the IRISA laboratory "Health-biology" cross-cutting axis [O. Dameron, until June 2019] [More details]
- Social commitee of Univ. Rennes 1 [C. Belleannée]
- Scientific committee of Univ. Rennes 1 school of medicine [O. Dameron, A. Siegel].
- Emergency aid commission of Univ. Rennes 1 & Rennes2 [C. Belleannée]

### *10.1.6. Research Administration*

*10.1.6.1. Inria Instances*

- Inria National evaluation board [A. Siegel, nominated member].
- Equality and diversity Committee, Inria - Responsible of the working group focusing on recruitment procedures [A. Siegel].

*10.1.6.2. CNRS*

- Chargée de mission "bioinformatique" at INS2I-CNRS (Institute for Computer Science of CNRS) [A. Siegel, nominated member].

*10.1.6.3. Inria local instances*

- Gender equality commission, IRISA & Inria Rennes [A. Siegel, coordinator]
- CUMI (Comission des utilisateurs des moyens informatiques) of Inria Rennes [F. Coste]
- Inria Rennes PhD recruitment (CORDIs) [C. Belleannée]

## 10.2. Teaching - Supervision - Juries

### *10.2.1. 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. Rennes1 [E. Becker, O. Dameron]

### *10.2.2. 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]
- "Atelier bioinformatique", Licence 2 informatique, Univ. Rennes 1 [O. Dameron]
- "Bioinformatique pour la génomique", 2nd year school of medicine, Univ. Rennes 1 [O. Dameron]
- "Bases de mathématiques et probablité", Master1 in public health, Univ. Rennes 1 [O. Dameron]
- "Programmation en Python", Master 1 in Public Health, Univ. Rennes 1 [O. Dameron]
- "Intégration: Remise à niveau en informatique", Master 1 in bioinformatics, Univ. Rennes 1 [O. Dameron]

- "Programmation impérative en Python", 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]
- Master: A. Siegel, Integrative and Systems biology, Master 2 in bioinformatics, Univ. Rennes 1 [A. Siegel]
- Micro-environnement Cellulaire normal & pathologique, Master 2 in Biologie cellulaire et Moléculaire, Univ. Rennes 1 [N. Théret]

### *10.2.3. Teaching*

Licence : E. Becker, "TPs Python", 12h, 1st year in Biology, Univ. Rennes 1, France

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", 43h, Master 1 in Bioinformatics, Univ. Rennes 1, France

Master : E. Becker, "Advanced R for data analysis", 34h, 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", 8h, 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

Licence: C. Belleannée, Langages formels, 20h, L3 informatique, Univ. Rennes1, France.

Licence: C. Belleannée, Projet professionnel et communication, 16h, L1 informatique, Univ. Rennes1, France.

Licence: C. Belleannée, Enseignant référent, 20h, L1 informatique, Univ. Rennes1, France.

Licence: C. Belleannée, Spécialité informatique : Functional and immutable programming , 42h, L1 informatique, Univ. Rennes1, France

Master: C. Belleannée, Algorithmique du texte et bioinformatique, 10h, M1 informatique, Univ. Rennes1, France

Master: S. Blanquart, Juries of Master 1 in bioinformatics, 5h, Univ. Rennes 1, France

Master: L. Bourneuf, Projet, 25h, M1 Santé Publique, France.

Master: F. Coste, Supervised machine learning, 20h, M2 + M1 Science Informatique, Univ. Rennes, France

Licence: O. Dameron, Biostatistiques, 12h, 1st year school of medicine, Univ. Rennes 1, France

Licence: O. Dameron, "Atelier bioinformatique", 12h, Licence 2 informatique, 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", 985h 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.

### *10.2.4. Supervision*

PhD: Yann Rivault, *Analyse de trajectoires de soins à partir de bases de données médico-administratives : apport d'un enrichissement pas des connaissances biomédicales issues du Web des Données*, defended 28th January 2019, supervised by O. Dameron and N. Lemeur [64].

PhD: Lucas Bourneuf, *A search space of graph motifs for graph compression: From Powergraphs to triplet concepts*, defended 17th december 2019, supervised by J. Nicolas [12].

PhD in progress: Johanne Bakalara, *Temporal models of care sequences for the exploration of medico-administrative data*, started in Oct. 2018, supervised by T. Guyet (Lacodam), E. Oger (Repères) and O. Dameron.

PhD in progress: Arnaud Belcour, *Inferring Model metabolisms for bacterial ecosystems reduction*, started in Oct. 2019, supervised by A. Siegel and S. Blanquart.

PhD in progress: 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: 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: 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: 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: Marine Louarn, *Intégration de données génomiques massives et hétérogènes, application aux mutations non-codantes dans le lymphome folliculaire*, started in Oct. 2017, supervised by A. Siegel, T. Fest (CHU) and O. Dameron.

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 : 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 : Méline Wéry, *Methodology development in disease treatment projects.* , started in Oct. 2017, supervised by E. Becker, C. Bettembourg (Sanofi), O. Dameron, and A. Siegel.

### *10.2.5. Juries*

- Referee of PhD thesis: Aarón Ayllón-Benítez, Univ. Bordeaux [O. Dameron] - A. Beica, ENS [A. Siegel, president] - F. Bridoux, Marseille [A. Siegel] - C. Hernandez, ENS [A. Siegel, president]
- Member of PhD thesis juries: Sandeep Manandhar, Univ. Rennes [O. Dameron] - A. Husson, Paris Diderot [A. Siegel] - Arran Hodgkinson, Univ. Montpellier [N. Théret]
- Referee of habilitation thesis: Fleur Mougin, Univ. Bordeaux [O. Dameron] - S. Peres, Orsay [A. Siegel]
- Member of habilitation thesis juries: Adrien Coulet, Univ. Lorraine [O. Dameron] - L. Calzone [A. Siegel, president]

### *10.2.6. Interns*

- Internship, from Jan 2019 until Jun 2019. Supervised by O. Dameron and O. Corby (Wimmics team, Sophia-Antipolis) . Student: Antoine Abel (Master 2 bioinformatique, Univ. Rennes 1). Subject: Faster SPARQL federated queries [31]
- Internship, from Jan 2019 until Jun 2019, Supervised by O. Dameron and A. Duchene (Theranexus). Student: Pierre Beaudier (Master 2 bioinformatique, Rennes). Subject: Evaluation of the public databases' relevance as comprehensive tools for pharmalogical mechanisms [32]
- Internship, from Jan 2019 until Jun 2019, Supervised by E. Becker, Fabrice Legeai (IGEPP) and Julien Bobe (LPGP). Student: Fanny Casse (Master 2 bioinformatique, Rennes). Subject: MiRNA regulating genes responsible for fertility and early embryonic development in the Medaka fish [35].
- Internship, from April 2019 until July 2019, Supervised by P. Dabert (IRSTEA), A. Siegel and S. Blanquart. Student: Theo Combe (Master 1 bioinformatique, Rennes). Subject: Reconstructing metabolic pathways of methanogenic communities using metabarcoding data
- Internship, from May 2019 until Jun 2019, Supervised by F. Coste. Student: Arthur Correnson (Licence math-informatique, Université Paul Sabatier, Toulouse). Subject: Implementation of a new partial local multiple alignment algorithm.
- Internship, from Jan 2019 until Jun 2019, Supervised by F. Coste, N. Théret, S. Blanquart, C. Belleannée. Student: Olivier Dennler (Master 2 bioinformatique, Rennes). Subject: Functional module characterization of ADAMTS / ADAMTSL protein family [37]
- Internship, from Jan 2019 until Jun 2019, Supervised by A. Siegel. Student: Clara Emery (Master 2 bioinformatique, Rennes). Subject: Modeling of quorum sensing system by in silico approaches
- Internship, from May 2019 until Oct 2019, Supervised by E. Becker and G. Rabut (IGDR, Rennes). Student: Camille Juigne (ENSSAT, Lannion). Subject: Data integration and formal analysis of the network between ubiquitination enzymes.
- Internship, from Jan 2019 until Jun 2019, Supervised by F. Morrews. Student: Albane Lysiak (Master 2 bioinformatique, Rennes). Subject: Exploiting paralog genes for discovering genetic regulation by transcription factors.
- Internship, from Jul 2019 until Aug 2019, Supervised by C. Belleannée and N. Guillaudeux. Student: Pierre Alexis Odye (Master 1 informatique, Rennes). Subject: Design of a database on known and predicted transcriptomics data in humans, mice and dogs.
- Internship, from Jul 2019 until Aug 2019, Supervised by E. Becker and F. Moreews. Student: Hugo Simon (1st year Telecom Paris Tech). Subject: Large scale analysis of metabolic networks as biological graphs.
- Internship, from Jun 2019 until Jul 2019, Supervised by A. Siegel and R. Andonov. Student: Kerian Thuillier (ENS Rennes). Subject: Linear programming for metabolic network completion [41]

- Internship, from Jan 2019 until Jul 2019, Supervised by C. Frioux. Student: Margot Wagner (Master 2 écologie, Rennes). Subject: Discriminant analysis of transcriptomics and metagenomics of an alga and its microbiota in multiple culture conditions with genome-scale modelling.

## 10.3. Popularization

### *10.3.1. Interventions*

- "J'peux pas j'ai informatique", Welcoming of high-school students to leverage stereotypes about computer sciences (110 participants), Apr. 2019 [O. Dameron, A. Siegel]
- "Elles codent Elles créent" Our Ph-D students have been involved in creative Python learning sessions for female students in two high-schools. [M. Louarn, M. Wéry, A. Siegel]
- Operation DECLIC (Dialogues Entre Chercheurs et Lycéens pour les Intéresser à la Construction des Savoirs). Lycée Descartes, Rennes. Nov 2019 [More details] [N. Théret, M. Conan, P. Vignet, O. Dennler]

### *10.3.2. Internal action*

- Organization of the weekly seminar "Symbiose", involving teams DYLISS, GENSCALE and GENOUEST: [S. Blanquart]

### *10.3.3. Creation of media or tools for science outreach*

**Science en Cour[t]s** (*http://sciences-en-courts.fr/*) Many of our current and former PhD students (M.Wéry, L. Bourneuf, A. Antoine-Lorquin, C. Bettembourg, J. Coquet, V. Delannée, G. Garet, S. Prigent) have been heavily involved in organization of a local Popularization Festival where PhD. students explain their thesis via short movies. The movies are presented to a professional jury composed of artists and scientists, and of high-school students. Previous years films can be viewed on the festival web-site [More details]

# 11. Bibliography

## Major publications by the team in recent years

[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, A. SIEGEL. *Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models*, in "PLoS Computational Biology", May 2018, vol. 14, n$^{o}$ 5, e1006146 [*DOI :* 10.1371/JOURNAL.PCBI.1006146], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01807842

[2] C. BELLEANNÉE, O. SALLOU, J. NICOLAS. *Logol: Expressive Pattern Matching in sequences. Application to Ribosomal Frameshift Modeling*, in "PRIB2014 - Pattern Recognition in Bioinformatics, 9th IAPR International Conference", Stockholm, Sweden, M. COMIN, L. KALL, E. MARCHIORI, A. NGOM, J. RAJAPAKSE (editors), Springer International Publishing, August 2014, vol. 8626, pp. 34-47 [*DOI :* 10.1007/978-3-319-09192-1_4], https://hal.inria.fr/hal-01059506

[3] C. BETTEMBOURG, C. DIOT, 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, 30 p. [*DOI :* 10.1371/JOURNAL.PONE.0133579], https://hal.inria.fr/hal-01184934

[4] P. BORDRON, M. LATORRE, M.-P. CORTÉS, M. GONZALES, S. THIELE, A. SIEGEL, A. MAASS, D. EVEILLARD. *Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach*, in "MicrobiologyOpen", 2015, vol. 5, n$^{o}$ 1, pp. 106-117 [*DOI :* 10.1002/MBO3.315], https://hal.inria.fr/hal-01246173

[5] J. COQUET, N. THÉRET, V. LEGAGNEUX, O. DAMERON. *Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering: Application to TGF-$\beta$ Signaling*, in "CMSB 2017 - 15th International Conference on Computational Methods in Systems Biology", Darmstadt, Lecture Notes in Computer Sciences, September 2017, 17 p. , https://hal.archives-ouvertes.fr/hal-01559249

[6] F. COSTE, G. GARET, A. GROISILLIER, J. NICOLAS, T. TONON. *Automated Enzyme classification by Formal Concept Analysis*, in "ICFCA - 12th International Conference on Formal Concept Analysis", Cluj-Napoca, Romania, Springer, June 2014, https://hal.inria.fr/hal-01063727

[7] F. COSTE, 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", Wrocław, Poland, September 2018, vol. 93, pp. 155 - 168, https://hal.inria.fr/hal-01872266

[8] C. FRIOUX, E. FREMY, C. TROTTIER, A. SIEGEL. *Scalable and exhaustive screening of metabolic functions carried out by microbial consortia*, in "Bioinformatics", September 2018, vol. 34, n$^{o}$ 17, pp. i934 - i943 [*DOI :* 10.1093/BIOINFORMATICS/BTY588], https://hal.inria.fr/hal-01871600

[9] C. FRIOUX, T. SCHAUB, S. SCHELLHORN, A. SIEGEL, P. WANKO. *Hybrid Metitebolic Network Completion*, in "Theory and Practice of Logic Programming", November 2018, pp. 1-23, 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, A. SIEGEL. *Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks*, in "PLoS Computational Biology", January 2017, vol. 13, n$^{o}$ 1, 32 p. [*DOI :* 10.1371/JOURNAL.PCBI.1005276], https://hal.inria.fr/hal-01449100

[11] S. VIDELA, J. SAEZ-RODRIGUEZ, C. GUZIOLOWSKI, A. SIEGEL. *caspo: a toolbox for automated reasoning on the response of logical signaling networks families*, in "Bioinformatics", 2017 [*DOI :* 10.1093/BIOINFORMATICS/BTW738], https://hal.inria.fr/hal-01426880

## Publications of the year

### Doctoral Dissertations and Habilitation Theses

[12] L. BOURNEUF. *A search space of graph motifs for graph compression : From Powergraphs to triplet concepts*, Université Rennes 1, December 2019, https://hal.inria.fr/tel-02399641

### Articles in International Peer-Reviewed Journals

[13] S. BLANQUART, A.-S. BOROWIEC, P. DELCOURT, M. FIGEAC, C. A. EMERLING, A. S. MESEGUER, M. ROUDBARAKI, N. PREVARSKAYA, G. BIDAUX. *Evolution of the human cold/menthol receptor, TRPM8*, in "Molecular Phylogenetics and Evolution", July 2019, vol. 136, pp. 104-118 [*DOI :* 10.1016/J.YMPEV.2019.04.011], https://hal.inria.fr/hal-02284919

[14] V. DELANNÉE, S. LANGOUËT, A. SIEGEL, N. THÉRET. *In silico prediction of Heterocyclic Aromatic Amines metabolism susceptible to form DNA adducts in humans*, in "Toxicology Letters", January 2019, vol. 300, pp. 18-30 [*DOI :* 10.1016/J.TOXLET.2018.10.011], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01903264

[15] W. DYRKA, M. PYZIK, F. COSTE, H. TALIBART. *Estimating probabilistic context-free grammars for proteins using contact map constraints*, in "PeerJ", March 2019, vol. 7, pp. 1-35 [*DOI :* 10.7717/PEERJ.6559], https://hal.archives-ouvertes.fr/hal-02400871

[16] M. FOLSCHETTE, V. LEGAGNEUX, A. PORET, L. CHEBOUBA, C. GUZIOLOWSKI, N. THÉRET. *A pipeline to create predictive functional networks: application to the tumor progression of hepatocellular carcinoma*, in "BMC Bioinformatics", January 2020 [*DOI :* 10.1186/s12859-019-3316-1], https://hal.archives-ouvertes.fr/hal-02095930

[17] D. NÈGRE, M. AITE, A. BELCOUR, C. FRIOUX, L. BRILLET-GUÉGUEN, X. LIU, P. BORDRON, O. GODFROY, A. P. LIPINSKA, C. LEBLANC, A. SIEGEL, S. M. DITTAMI, E. CORRE, G. V. MARKOV. *Genome–Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae Saccharina japonica and Cladosiphon okamuranus*, in "Antioxidants", November 2019, vol. 8, n$^{o}$ 11, 564 p. [*DOI :* 10.3390/ANTIOX8110564], https://hal.inria.fr/hal-02395080

[18] Y. RIVAULT, O. DAMERON, N. LE MEUR. *queryMed: Semantic Web functions for linking pharmacological and medical knowledge to data*, in "Bioinformatics", January 2019, vol. 35, n$^{o}$ 17, pp. 3203-3205 [*DOI :* 10.1093/BIOINFORMATICS/BTZ034], https://hal.archives-ouvertes.fr/hal-01988699

[19] K. STRAUB, M. LINDE, C. KROPP, S. BLANQUART, P. BABINGER, R. MERKL. *Sequence selection by FitSS4ASR alleviates ancestral sequence reconstruction as exemplified for geranylgeranylglyceryl phosphate synthase*, in "Biological Chemistry", February 2019, vol. 400, n$^{o}$ 3, pp. 367-381 [*DOI :* 10.1515/HSZ-2018-0344], https://hal.inria.fr/hal-02284913

[20] M. WERY, O. DAMERON, J. NICOLAS, E. RÉMY, A. SIEGEL. *Formalizing and enriching phenotype signatures using Boolean networks*, in "Journal of Theoretical Biology", 2019, vol. 467, pp. 66-79 [*DOI :* 10.1016/J.JTBI.2019.01.015], https://hal.inria.fr/hal-02018724

### Invited Conferences

[21] A. SIEGEL. *Learning boolean regulations based on prior-knowledge data: a logical-based viewpoint*, in "SASB 2019 - 10th International Workshop on Static Analysis for Systems Biology", Porto, Portugal, October 2019, https://hal.inria.fr/hal-02412421

[22] A. SIEGEL. *Using automated reasoning to explore unconventional organisms: a first step to explore host-microbial interactions*, in "MPA 2019 - Conference on Metabolic Pathway Analysis", Riga, Latvia, August 2019, https://hal.inria.fr/hal-02412419

### International Conferences with Proceedings

[23] J. BAKALARA, T. GUYET, O. DAMERON, E. OGER, A. HAPPE. *Temporal models of care sequences for the exploration of medico-administrative data*, in "2019 - Workshop IA&Santé, PFIA", Toulouse, France, July 2019, pp. 1-8, https://hal.archives-ouvertes.fr/hal-02265743

[24] L. BOURNEUF, J. NICOLAS. *Concept Lattices as a Search Space for Graph Compression*, in "ICFCA 2019 - 15th International Conference on Formal Concept Analysis", Francfort, Germany, D. C. L. B. SERTKAYA (editor), ICFCA: International Conference on Formal Concept Analysis, Springer, May 2019, vol. 15th International Conference, n$^{o}$ 15, pp. 274-289 [*DOI :* 10.1007/978-3-030-21462-3_18], https://hal.inria.fr/hal-02399578

[25] M. LOUARN, F. CHATONNET, X. GARNIER, T. FEST, A. SIEGEL, O. DAMERON. *Increasing life science resources re-usability using Semantic Web technologies*, in "eScience 2019 - 15th International eScience Conference", San Diego, United States, September 2019, pp. 1-9, https://hal.inria.fr/hal-02274982

### Conferences without Proceedings

[26] L. BOURNEUF. *Biseau: An Answer Set Programming Environment for High-Level Specification and Graph Visualization applied to FCA*, in "ICFCA 2019 - International Conference on Formal Concept Analysis", Frankfort, Germany, June 2019, pp. 1-6, https://hal.inria.fr/hal-02399610

[27] X. GARNIER, A. BRETAUDEAU, F. LEGEAI, A. SIEGEL, O. DAMERON. *AskOmics: a user-friendly interface to Semantic Web technologies for integrating local datasets with reference resources*, in "JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques", Nantes, France, July 2019, 1 p. , https://hal.archives-ouvertes.fr/hal-02401750

[28] V. J. HENRY, G. BASSIGNANA, V. ZUJOVIC, F. DE VICO FALLANI, O. DAMERON, I. MOSZER, O. COLLIOT. *Conciliation of process description and molecular interaction networks using logical properties of ontology*, in "JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques", Nantes, France, July 2019, https://hal.archives-ouvertes.fr/hal-02301702

[29] N. LE MEUR, Y. RIVAULT, O. DAMERON. *Linking pharmacological and medical knowledge using semantic Web technologies*, in "useR! 2019 - Conférence internationale des utilisateurs de R", Toulouse, France, July 2019, https://hal.archives-ouvertes.fr/hal-02294364

[30] H. TALIBART, F. COSTE. *Using residues coevolution to search for protein homologs through alignment of Potts models*, in "CECAM 2019 - workshop on Co-evolutionary methods for the prediction and design of protein structure and interactions", Lausanne, Switzerland, CECAM-HQ-EPFL, June 2019, pp. 1-2, https://hal.inria.fr/hal-02402646

### Other Publications

[31] A. ABEL. *Faster SPARQL Federated Queries*, Université Rennes1, June 2019, https://hal.inria.fr/hal-02269417

[32] P. BEAUDIER. *Evaluation of the public databases' relevance as comprehensive tools for pharmacological mechanisms*, Université de Rennes 1, June 2019, https://hal.inria.fr/hal-02191210

[33] A. BELCOUR, C. FRIOUX, M. AITE, A. BRETAUDEAU, A. SIEGEL. *Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species*, December 2019, working paper or preprint [*DOI :* 10.1101/803056], https://hal.inria.fr/hal-02395024

[34] B. BURGUNTER-DELAMARE, H. KLEINJAN, C. FRIOUX, E. FREMY, M. WAGNER, E. CORRE, A. LE SALVER, C. LEROUX, C. LEBLANC, A. SIEGEL, S. M. DITTAMI. *Metabolic complementarity between a brown alga and associated 1 cultivable bacteria provide indications of beneficial interactions*, October 2019, working paper or preprint [*DOI :* 10.1101/813683], https://hal.sorbonne-universite.fr/hal-02333039

[35] F. CASSE ENCADRÉE. *Regulation of genes regulating fecondity and early embryonic development by miRNAs in the Medaka fish*, Univ Rennes, Inria, CNRS, IRISA, June 2019, https://hal.inria.fr/hal-02192476

[36] F. COSTE. *Deep learning languages: a key fundamental shift from probabilities to weights?*, August 2019, https://arxiv.org/abs/1908.00785 - working paper or preprint, https://hal.inria.fr/hal-02235207

[37] O. DENNLER. *Caractérisation en modules fonctionnels de la famille de protéines ADAMTS / ADAMTSL*, Univ Rennes, June 2019, https://hal.inria.fr/hal-02403084

[38] S. M. DITTAMI, E. CORRE, L. BRILLET-GUÉGUEN, A. P. LIPINSKA, 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. HERVÉ, D. JOLLIVET, H. KLEINJAN, C. LEBLANC, X. LIU, D. MARIE, G. V. MARKOV, A. E. MINOCHE, P. PERICARD, M.-M. PERRINEAU, A. F. PETERS, A. SIEGEL, C. TROTTIER, H. S. YOON, H. HIMMELBAUER, C. BOYEN, T. TONON. *The genome of Ectocarpus subulatus -a highly stress-tolerant brown alga*, October 2019, working paper or preprint [*DOI :* 10.1101/307165], https://hal.sorbonne-universite.fr/hal-02333021

[39] N. GUILLAUDEUX, C. BELLEANNÉE, S. BLANQUART, J.-S. VARRÉ. *Predicting isoform transcripts: What does the comparison of known transcripts in human, mouse and dog tell us?*, July 2019, vol. 8, 1 p. , JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster présenté à JOBIM 2019 (https://jobim2019.sciencesconf.org/) qui s'est déroulé du 2 au 5 juillet 2019 et à ISMB/ECCB 2019 (https://www.iscb.org/ismbeccb2019) qui s'est déroulé du 21 au 25 juillet 2019 [*DOI :* 10.7490/F1000RESEARCH.1117311.1], https://hal.inria.fr/hal-02267357

[40] H. TALIBART, F. COSTE. *Using residues coevolution to search for protein homologs through alignment of Potts models*, July 2019, 1 p. , JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster, https://hal.inria.fr/hal-02402687

[41] K. THUILLIER. *Linear programming for metabolic network completion*, Inria Rennes - Bretagne Atlantique and University of Rennes 1, France, December 2019, https://hal.inria.fr/hal-02408003

[42] M. WERY, E. BECKER, F. AUGE, C. BETTEMBOURG, O. DAMERON, A. SIEGEL. *Identification of causal signature using omics data integration and network reasoning-based analysis*, July 2019, 1 p. , JOBIM 2019 - Journées Ouvertes Biologie, Informatique et Mathématiques, Poster, https://hal.inria.fr/hal-02193860

## References in notes

[43] G. ANDRIEUX, M. LE BORGNE, N. THÉRET. *An integrative modeling framework reveals plasticity of TGF-Beta signaling*, in "BMC Systems Biology", 2014, vol. 8, n$^{o}$ 1, 30 p. [*DOI :* 10.1186/1752-0509-8-30], http://www.hal.inserm.fr/inserm-00978313

[44] T. BERNERS LEE, W. HALL, J. A. HENDLER, K. O'HARA, N. SHADBOLT, D. J. WEITZNER. *A Framework for Web Science*, in "Foundations and Trends in Web Science", 2007, vol. 1, n$^{o}$ 1, pp. 1–130

[45] C. BETTEMBOURG, C. DIOT, O. DAMERON. *Semantic particularity measure for functional characterization of gene sets using gene ontology*, in "PLoS ONE", 2014, vol. 9, n$^{o}$ 1, e86525 [*DOI :* 10.1371/JOURNAL.PONE.0086525], https://hal.inria.fr/hal-00941850

[46] S. BLANQUART, J.-S. VARRÉ, P. GUERTIN, A. PERRIN, A. BERGERON, K. M. SWENSON. *Assisted transcriptome reconstruction and splicing orthology*, in "BMC Genomics", Nov 2016, vol. 17, n$^{o}$ 10, 786 p. , https://doi.org/10.1186/s12864-016-3103-6

[47] P. BLAVY, F. GONDRET, S. LAGARRIGUE, J. VAN MILGEN, 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", 2014, vol. 8, n$^{o}$ 1, 32 p. [*DOI :* 10.1186/1752-0509-8-32], https://hal.inria.fr/hal-00980499

[48] P. BORDRON, M. LATORRE, M.-P. CORTÉS, M. GONZALES, S. THIELE, A. SIEGEL, A. MAASS, D. EVEILLARD. *Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach*, in "MicrobiologyOpen", 2015, vol. 5, n$^{o}$ 1, pp. 106-117 [*DOI :* 10.1002/MBO3.315], https://hal.inria.fr/hal-01246173

[49] M. BOUTET, L. GAUTHIER, M. LECLERC, G. GROS, V. DE MONTPREVILLE, N. THÉRET, E. DONNADIEU, F. MAMI-CHOUAIB. *TGF-$\beta$ signaling intersects with CD103 integrin signaling to promote T lymphocyte accumulation and antitumor activity in the lung tumor microenvironment*, in "Cancer Research", 2016, vol. 76, n$^{o}$ 7, pp. 1757-69 [*DOI :* 10.1158/0008-5472.CAN-15-1545], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01282442

[50] A. BRETAUDEAU, F. COSTE, F. HUMILY, L. GARCZAREK, G. LE CORGUILLÉ, C. SIX, M. RATIN, O. COLLIN, W. M. SCHLUCHTER, F. PARTENSKY. *CyanoLyase: a database of phycobilin lyase sequences, motifs and functions*, in "Nucleic Acids Research", November 2012, 6 p. [*DOI :* 10.1093/NAR/GKS1091], https://hal.inria.fr/hal-01094087

[51] J. COQUET, N. THÉRET, V. LEGAGNEUX, O. DAMERON. *Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering: Application to TGF-$\beta$ Signaling*, in "CMSB 2017 - 15th International Conference on Computational Methods in Systems Biology", Darmstadt, France, Lecture Notes in Computer Sciences, September 2017, 17 p. , https://hal.archives-ouvertes.fr/hal-01559249

[52] M.-P. CORTÉS, S. N. MENDOZA, D. TRAVISANY, A. GAETE, A. SIEGEL, V. CAMBIAZO, A. MAASS. *Analysis of Piscirickettsia salmonis Metabolism Using Genome-Scale Reconstruction, Modeling, and Testing*, in "Frontiers in Microbiology", December 2017, vol. 8, 15 p. [*DOI :* 10.3389/FMICB.2017.02462], https://hal.inria.fr/hal-01661270

[53] F. COSTE, G. GARET, A. GROISILLIER, J. NICOLAS, T. TONON. *Automated Enzyme classification by Formal Concept Analysis*, in "ICFCA - 12th International Conference on Formal Concept Analysis", Cluj-Napoca, Romania, Springer, June 2014, https://hal.inria.fr/hal-01063727

[54] V. DELANNÉE, S. LANGOUËT, N. THÉRET, A. SIEGEL. *A modeling approach to evaluate the balance between bioactivation and detoxification of MeIQx in human hepatocytes*, in "PeerJ", 2017, vol. 5, e3703 p. [*DOI :* 10.7717/PEERJ.3703], https://hal.inria.fr/hal-01575579

[55] 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, 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", 2014, vol. 5, 241 p. [*DOI :* 10.3389/FGENE.2014.00241], https://hal.inria.fr/hal-01079739

[56] K. FAUST, J. RAES. *Microbial interactions: from networks to models*, in "Nat. Rev. Microbiol.", Jul 2012, vol. 10, n$^{o}$ 8, pp. 538–550

[57] M. Y. GALPERIN, D. J. RIGDEN, X. M. FERNÁNDEZ-SUÁREZ. *The 2015 nucleic acids research database issue and molecular biology database collection*, in "Nucleic acids research", 2015, vol. 43, n$^{o}$ D1, pp. D1–D5

[58] M. GEBSER, R. KAMINSKI, B. KAUFMANN, T. SCHAUB. *Answer Set Solving in Practice*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan and Claypool Publishers, 2012

[59] F. GONDRET, I. LOUVEAU, M. HOUEE, D. CAUSEUR, A. SIEGEL. *Data integration*, in "Meeting INRA-ISU", Ames, United States, March 2015, 11 p. , https://hal.archives-ouvertes.fr/hal-01210940

[60] F. HERAULT, A. VINCENT, O. DAMERON, P. LE ROY, P. CHEREL, M. DAMON. *The longissimus and semimembranosus muscles display marked differences in their gene expression profiles in pig*, in "PLoS ONE", 2014, vol. 9, n$^{o}$ 5, e96491 [*DOI :* 10.1371/JOURNAL.PONE.0096491], https://hal.inria.fr/hal-00989635

[61] M.-A. LAURENT, D. BONNIER, N. THÉRET, P. TUFFÉRY, G. MOROY. *In silico characterization of the interaction between LSKL peptide, a LAP-TGF-beta derived peptide, and ADAMTS1*, in "Computational Biology and Chemistry", April 2016, vol. 61, pp. 155-161 [*DOI :* 10.1016/J.COMPBIOLCHEM.2016.01.012], https://hal.archives-ouvertes.fr/hal-02394687

[62] S. PRIGENT, G. COLLET, S. M. DITTAMI, L. DELAGE, F. ETHIS DE CORNY, O. DAMERON, D. EVEILLARD, S. THIELE, J. CAMBEFORT, C. BOYEN, A. SIEGEL, T. TONON. *The genome-scale metabolic network of Ectocarpus siliculosus (EctoGEM): a resource to study brown algal physiology and beyond*, in "Plant Journal", September 2014, pp. 367-81 [*DOI :* 10.1111/TPJ.12627], https://hal.archives-ouvertes.fr/hal-01057153

[63] S. PRIGENT, C. FRIOUX, S. M. DITTAMI, S. THIELE, A. LARHLIMI, G. COLLET, G. FABIEN, J. GOT, D. EVEILLARD, J. BOURDON, F. PLEWNIAK, T. TONON, A. SIEGEL. *Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks*, in "PLoS Computational Biology", January 2017, vol. 13, n$^{o}$ 1, 32 p. [*DOI :* 10.1371/JOURNAL.PCBI.1005276], https://hal.inria.fr/hal-01449100

[64] Y. RIVAULT. *Care trajectory analysis using medico-administrative data : contribution of a knowledge-based enrichment from the Linked Data*, Université Rennes 1, January 2019, https://tel.archives-ouvertes.fr/tel-02137442

[65] M. H. SAIER, V. S. REDDY, B. V. TSU, M. S. AHMED, C. LI, G. MORENO-HAGELSIEB. *The Transporter Classification Database (TCDB): recent advances*, in "Nucleic Acids Res.", Jan 2016, vol. 44, n$^{o}$ D1, pp. D372–379

[66] D. B. SEARLS. *String variable grammar: A logic grammar formalism for the biological language of DNA*, in "The Journal of Logic Programming", 1995, vol. 24, n$^{o}$ 1, pp. 73 - 102, Computational Linguistics and Logic Programming [*DOI :* 10.1016/0743-1066(95)00034-H], http://www.sciencedirect.com/science/article/pii/074310669500034H

[67] Z. D. STEPHENS, S. Y. LEE, F. FAGHRI, R. H. CAMPBELL, C. ZHAI, M. J. EFRON, R. IYER, M. C. SCHATZ, S. SINHA, G. E. ROBINSON. *Big Data: Astronomical or Genomical?*, in "PLoS biology", 2015, vol. 13, n$^{o}$ 7, e1002195 p.

[68] R. VERBORGH, M. VANDER SANDE, O. HARTIG, J. VAN HERWEGEN, L. DE VOCHT, B. DE MEESTER, G. HAESENDONCK, P. COLPAERT. *Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web*, in "Journal of Web Semantics", March 2016, vol. 37–38, pp. 184–206 [*DOI :* DOI:10.1016/J.WEBSEM.2016.03.003], http://linkeddatafragments.org/publications/jws2016.pdf