UMR IRISA

Activity Report 2022

Team DYLISS

Dynamics, Logics and Inference for biological Systems and Sequences

Joint team with Centre Inria de l'Université de Rennes

D7 – Data and Knowledge Management

Université de Rennes | rennes | INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES | cnrs | Inria | Inserm | CentraleSupélec | IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom | Université Bretagne Sud ubs

Contents

Project-Team DYLISS

Creation of the Project-Team: 2013 July 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.7. – Open data

A3.1.10. – Heterogeneous data

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.2.6. – Linked data

A3.3.3. – Big data analysis

A7.2. – Logic in Computer Science

A8.1. – Discrete mathematics, combinatorics

A8.2. – Optimization

A9.1. – Knowledge

A9.2. – Machine learning

A9.7. – AI algorithmics

A9.8. – Reasoning

Other research topics and application domains

B1.1.2. – Molecular and cellular biology

B1.1.7. – Bioinformatics

B1.1.10. – Systems and synthetic biology

B2.2.3. – Cancer

B2.2.5. – Immune system diseases

1 Team members, visitors, external collaborators

Research Scientists

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

Faculty Members

- Olivier Dameron [Team leader, UNIV RENNES I, Professor, HDR]
- Emmanuelle Becker [UNIV RENNES I, Associate Professor, HDR]
- Catherine Belleannée [UNIV RENNES I, Associate Professor]
- Yann Le Cunff [UNIV RENNES I, Associate Professor]

Post-Doctoral Fellows

- Arnaud Belcour [UNIV RENNES (oct.) and CNRS ANR SeaBioz (dec.), from Oct 2022]
- Victor Mataigne [CNRS, from Sep 2022]

PhD Students

- Arnaud Belcour [INRIA, until Sep 2022]
- Matthieu Bougueon [INSERM]
- Nicolas Buton [UNIV RENNES I]
- Olivier Dennler [INSERM, until Sep 2022]
- Camille Juigne [INRAE]
- Virgilio Kmetzsch Rosa E Silva [INRIA, until Sep 2022]
- Marc Melkonian [Centre Hospitalier de Centre Bretagne, until Nov 2022]
- Baptiste Ruiz [INRIA]
- Kerian Thuillier [CNRS]

Technical Staff

- Olivier Dennler [INRIA (oct.) and UNIV RENNES (nov. to dec.), Engineer, from Oct 2022]
- Jeanne Got [CNRS, Engineer]
- Pauline Hamon-Giraud [INRIA , Engineer, from Apr 2022 until Jul 2022]
- François Moreews [INRAE, Engineer]
- Yael Tirlet [CNRS ANR DeepImpact, Engineer, from Nov 2022]

Interns and Apprentices

- Thibaut Antoine [UNIV RENNES I , Intern, from May 2022 until Jul 2022]
- Moana Aulagner [INRIA , Intern, from Apr 2022 until Jun 2022]
- Cécile Beust [INRIA, Intern, from Apr 2022 until Jul 2022]
- Pauline Hamon-Giraud [CNRS, Intern, until Jul 2022, Engineer since Nov. 2022]
- Yael Tirlet [INRAE, until Jul 2022, Intern, Engineer since Oct. 2022]

Administrative Assistant

- Marie Le Roic [INRIA]

Visiting Scientists

- Oumarou Abdou-Arbi [UNIV UDDM MARADI, from Aug 2022 until Sep 2022]
- Alejandro Maass [UAI SANTIAGO CHILI, from Oct 2022]

External Collaborators

- Denis Tagu [INRAE]
- Nathalie Theret [INSERM, HDR]

2 Overall objectives

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

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

Computational challenges. The main challenges we face are **data incompleteness and heterogeneity, leading to non-identifiability**. Indeed, we have observed that the biological systems that we consider cannot be uniquely identifiable. Indeed, "omics" technologies have allowed the number of measured

compounds in a system to increase tremendously. However, it appears that the theoretical number of different experimental measurements required to integrate these compounds in a single discriminative model has increased exponentially with respect to the number of measured compounds. Therefore, according to the current state of knowledge, there is no possibility to explain the data with a single model. Our rationale is that biological systems will still remain non-identifiable for a very long time. In this context, we favor **the construction and the study of a space of feasible models or hypotheses**, including known constraints and facts on a living system, rather than searching for a single discriminative optimized model. We develop methods allowing a precise and exhaustive investigation of this space of hypotheses. With this strategy, we are in the position of developing experimental strategies to progressively shrink the space of hypotheses and increase the understanding of the system.

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

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

3 Research program

3.1 Context: Computer science perspective on symbolic artificial intelligence

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

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

Reasoning over structured data with constraint-based logical paradigms Another level of complexity in life science integration is that very few paradigms exist to model the behavior of a complex biological system. This leads biologists to perform and formulate hypotheses in order to interpret their data. Our strategy is to interpret such hypotheses as combinatorial optimization problems, allowing to reduce the family of models compatible with data. To that goal, we collaborate with Potsdam University in order to use and challenge the most recent developments of Answer Set Programming (ASP) [69], a logical paradigm for solving constraint satisfiability and combinatorial optimization issues.

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

Characterizing biological sequences with formal syntactic models Our last goal is to identify and characterize the function of expressed genes such as transcripts, enzymes or isoforms in non-model species biological networks or specific functional features of metagenomic samples. These are insufficiently

precise because of the divergence of biological sequences, the complexity of molecular structures and biological processes, and the weak signals characterizing these elements.

Our goal is therefore to develop accurate formal syntactic models (automata, grammars or abstract gene models) that would enable us to represent sequence conservation, sets of short and degenerated patterns, and crossing or distant dependencies. This requires both to determine the classes of formal syntactic models adequate for handling biological complexity, and to automatically characterize the functional potential embodied in biological sequences with these models.

3.2 Scalable methods to query data heterogenity

Confronted to large and complex data sets (raw data are associated with graphs depicting explicit or implicit links and correlations) almost all scientific fields have been impacted by the *big data issue*, especially genomics and astronomy [80]. In our opinion, life sciences cumulate several features that are very specific and prevent the direct application of big data strategies that proved successful in other domains such as experimental physics: the existence of **several scales of granularity** (from microscopic to macroscopic) and the associated issue of dependency propagation, datasets **incompleteness and uncertainty** (including highly **heterogeneous** responses to a perturbation from one sample to another), and highly fragmented sources of information that **lacks interoperability** [67]. To explore this research field, we use techniques from symbolic data mining (Semantic Web technologies, symbolic clustering, constraint satisfaction, and grammatical modeling) to take into account those life science features in the analysis of biological data.

3.2.1 Research topics

Facilitating data integration and querying The quantity and inner complexity of life science data require semantically-rich analysis methods. A major challenge is then to combine data (from local project as well as from reference databases) and symbolic knowledge seamlessly. Semantic Web technologies (RDF for annotating data, OWL for representing symbolic knowledge, and SPARQL for querying) provide a relevant framework, as demonstrated by the success of Linked (Open) Data [51]. 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* [86], query properties and dataset structure for decomposing federated SPARQL queries.

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

3.2.2 Associated software tools

AskOmics platform AskOmics is an integration and interrogation software for linked biological data based on semantic web technologies[1]. AskOmics aims at bridging the gap between end user data and the Linked (Open) Data cloud (LOD cloud). It allows heterogeneous bioinformatics data (formatted as tabular files or directly in RDF) to be loaded into a Triple Store system using a user-friendly web interface. It helps end users (1) to take advantage of the information available in the LOD cloud for analyzing there own data, and (2) to contribute back to the linked data by representing their data and the associated metadata

[1] askomics.org

in the proper format, as well as by linking them to other resources. An originality is the graphical interface that allows any dataset to be integrated in a local RDF datawarehouse and SPARQL query to be built transparently and iteratively by a non-expert user.
Pax2graphml aims at easily manipulating BioPAX source files as regulated reaction graphs described in graph format. The goal is to be highly flexible and to integrate graphs of regulated reactions from a single BioPAX source or by combining and filtering BioPAX sources. The output graphs can then be analyzed with additional tools developed in the team, such as KeyRegulatorFinder.

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

3.3 Metabolism: from protein sequences to systems ecology

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

3.3.1 Research topics

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

System level: enriching and comparing metabolic networks for non-model organisms
Non-model organisms often lack both complete and reliable annotated sequences, which cause the draft networks of their metabolism to largely suffer from incompleteness. In former studies, the team has developed several methods to improve the quality of eukaryotic metabolic networks, by solving several variants of the so-called *Metabolic Network gap-filling problem* with logical programming approaches [9, 8]. The main drawback of these approaches is that they cannot scale to the reconstruction and comparison of families of metabolic networks. Our main objective is therefore to develop new tools for the comparison of species strains at the metabolic level.

Consortium level: exploring the diversity of community consortia The newly emerging field of system ecology aims at building predictive models of species interactions within an ecosystem, with the goal of deciphering cooperative and competitive relationships between species [66]. This field raises two new issues: (1) uncertainty on the species present in the ecosystem and (2) uncertainty about the global objective governing an ecosystem. To address these challenges, our first research focus is the inference of metabolic exchanges and relationships for transporter identification, based on our expertise in metabolic network gap-filling. The second challenging focus is the prediction of transporters families via refined characterization of transporters, which are quite unexplored apart from specific databases [78].

3.3.2 Associated software tools

Protomata[5] is a machine learning suite for the inference of automata characterizing (functional) families of proteins at the sequence level. It provides programs to build a new kind of sequence alignments

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

(characterized as partial and local), learn automata, and search for new family members in sequence databases. By enabling to model local dependencies between positions, automata are more expressive than classical tools (PSSMs, Profile HMMs, or Prosite Patterns) and are well suited to predict new family members with a high specificity. This suite is for instance embedded in the cyanolase database [57] to automate its updade and was used for refining the classification of HAD enzymes [6] or identify shared conservations in the core proteome of extracellular vesicles produced by human and animal *S. aureus* strains [83].

PPSuite[6] is one of the first frameworks taking into account coevolutionary dependencies between residues for the comparison of protein sequences. It proposes a complete workflow enabling to infer direct couplings between the positions of a sequence of interest by a Potts model with the help of the sequence close homologs and to score the similarity of the sequences by alignment of the inferred Potts models, as well as tools to visualize the models and their alignments [82, 81].

AuReMe and AuCoMe workspaces is designed for tractable reconstruction of metabolic networks[7]. The toolbox allows for the Automatic Reconstruction of Metabolic networks based on the combination of multiple heterogeneous data and knowledge sources [1]. The main added values are the inclusion of graph-based tools relevant for the study of non-model organisms (Meneco and Menetools packages), the possibility to trace the reconstruction and curation procedures (Padmet package), and the exploration of reconstructed metabolic networks with wikis (wiki-export package, see: aureme.genouest.org/wiki.html). It also generates outputs to explore the resulting networks with Askomics. It has been used for reconstructing metabolic networks of micro and macro-algae [76], extremophile bacteria [60] and communities of organisms [4].

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

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

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

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

3.4.1 Research topics

Genomic level: characterizing gene structure with grammatical languages and conservation information The goal here is to accurately represent gene structure, including intron/exon structure, for predicting the products of genes, such as isoform transcripts, and comparing the expression potential of a eukaryotic gene according to its context (e.g. tissue) or according to the species. Our approach consists in designing grammatical and comparative-genomics based models for gene structures able

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

to detect heterogeneous functional sites (splicing sites, regulatory binding sites...), functional regions (exons, promotors...) and global constraints (translation into proteins) [53]. 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 [79].

System level: extracting causal signatures of complex phenotypes with systems biology frameworks Our main challenge is to set up a generic formalism to model inter-layer interactions in large-scale biological networks. To that goal, we have developed several types of abstractions: multi-experiments framework to learn and control signaling networks [10], multi-layer reactions in interaction graphs [54], and multi-layer information in large-scale Petri nets [49]. Our main issues are to scale these approaches to standardized large-scale repositories by relying on the interoperable Linked Open Data (LOD) resources and to enrich them with ad-hoc regulations extracted from sequence-based analysis. This will allow us to characterize changes in system attractors induced by mutations and how they may be included in pathology signatures.

3.4.2 Associated software tools

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

Caspo Cell ASP Optimizer (Caspo) software constitutes a pipeline for automated reasoning on logical signaling networks (learning, classifying, designing experimental perturbations, identifying controllers, take time-series into account)[11]. The software handles inherent experimental noise by enumerating all different logical networks which are compatible with a set of experimental observations [10]. The main advantage is that it enables a complete study of logical network without requiring any linear constraint programs.

Cadbiom package aims at building and analyzing the asynchronous dynamics of enriched logical networks[12]. It is based on Guarded transition semantic and allows synchronization events to be investigated in large-scale biological networks [49]. For example, it allowed to analyze controler of phenotypes in a large-scale knowledge database (PID) [5].

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

4 Application domains

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

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

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

- **Protein functions in seaweed metabolism** Several years ago, our methods based on combinatorial optimization for the reconstruction of genome-scale metabolic networks and on classification of enzyme families based on local and partial alignments allowed the seaweed *E. siliculosus* metabolism to be deciphered [76, 61]. The study of the *HAD* superfamily of proteins thanks to partial local alignments produced by Protomata tools, allowed sub-families to be deciphered and classified. Additionally, the metabolic map reconstructed with Meneco enabled the reannotation of 56 genes within the *E. siliculosus* genome. These approaches also shed light on evolution of metabolic processes.

- **Elucidating algal metabolism thanks to large-scale metabolic network reconstructions** More recently, the tools developed by Dyliss (based on the AuReMe toolbox) allowed us to participate in the reconstruction of a metabolic network for the brown algae *Saccharina japonica* and *Cladosiphon okamuranus* in order to identify these species specificities on the synthesis of carotenoids biosynthesis [75]. We also participated in the study of the genome of *Ectocarpus subulatus*, a highly stress-tolerant algal strain [65]. Finally, AuReMe has been used to analyze the metabolic capacity of several strains of cyanobacteria, with results integrated in the Cyanorak database [68] and to characterize synergistic effects of the *synechococcus* strain WH7803 [71].

- **Metabolic pathway drift theory** Genome annotations can contribute to understanding algal metabolism. The tool PathModel was developed to add support for biochemical reactions and metabolite structures to the theory of metabolic pathway drift with an approach combining chemoinformatics knowledge reasoning and modeling. This approach was applied to the study of the red alga *Chondrus crispus*, which allowed to show that even for metabolic pathways supposed to be conserved between species (sterols, mycrosporins synthesis), we can see an important turnover in the order of reactions appearing in a metabolic pathway. This work lays the foundations for the concept of "metabolic drift" analogous to the same concept in genomics. [50].

- **Algal-bacteria interactions** We reconstructed the metabolic network of a symbiot bacterium *Ca. P. ectocarpi* [64] and used this reconstructed network to decipher interactions within the algal-bacteria holobiont, revealing several candidates metabolic pathways for algal-bacterial interactions. Similarily, our analyses suggested that the bacterium *Ca. P. ectocarpi* is able to provide both beta-alanine and vitamin B5 to the seaweed via the phosphopantothenate biosynthesis pathway [77].

  These works paved the way to the study of host-microbial interactions, as shown in [58] where we evidenced the role of tools such as miscoto and metage2metabo to predict synthetic communities allowing to restore algal metabolic pathways. To validate these approaches experimentally, we worked with S. Dittami, researcher at the Roscoff biological station. We applied these methods on a set of about fifteen cultivable bacteria identified on the wall membrane of *Ectocarpus siliculosus*. Our approaches predicted that three bacteria were necessary to facilitate the growth of this alga in an axenic medium. The experiments were carried out, and indeed allowed the alga to grow in an axenic medium. This is therefore a proof of concept of the relevance of our approaches. More recently, the study of the freshwater strain of *Ectocarpus subulatus* evidenced the role of metabolism in adaptation, paving the way to biotechnological applications [13].

Microbiology: elucidating the functioning of extremophile consortiums of bacteria. Our main issue is the understanding of bacteria living in extreme environments. The context is mainly a collaboration with the group of bioinformatics at Universidad de Chile (co-funded by the Center of Mathematical Modeling, the Center of Regulation Genomics and Inria-Chile). In order to elucidate the main characteristics of these bacteria, our integrative methods were developed to identify the main groups of regulators for their specific response in their living environment. The integrative biology tools Meneco, Lombarde and Shogen have been designed in this context. In particular, genome-scale metabolic network been recently reconstructed and studied with the Meneco and Shogen approaches, especially on bacteria involved in biomining processes [55] and in Salmon pathogenicity [60]. We have also studied the specificities of two Microbacterium strains, CGR1 and CGR2, isolated in different soils of the Atacama Desert in Chile,

showing significant differences on the connectivity of metabolite production in relation to pH tolerance and CO2 production [74].

Agriculture and environmental sciences: upstream controllers of cow, pork and pea-aphid metabolism and regulation. Our goal is to propose methods to identify regulators of complex phenotypes related to environmental issues. Our work on the identification of upstream regulators within large-scale knowledge databases (tool KeyRegulatorFinder) [54] and on semantic-based analysis of metabolic networks [52] was very valuable for interpreting the differences of gene expression in pork meat [72] and figure out the main gene-regulators of the response of porks to several diets [70]. Our expertise in microbiota analysis is also currently being applied to rumen microbial genomics [20].

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

- *TGF-β signaling networks.* TGF-β is a multifunctional cytokine that binds to specific receptors and induce numerous signaling pathways depending on the context. Deciphering TGF-β signaling networks requires to take into account a system-wide view and develop predictive models for therapeutic benefit. For that purpose we developed Cadbiom and identified gene networks associated with innate immune response to viral infection that combine TGF-β and interleukin signaling pathways [49, 59]. More recently we have very significantly refactored Cadbiom package towards a framework that allows the identification of causal regulators in large-scale models, formalized in the BioPAX language and automatically interpreted as guarded transitions[13].The Cadbiom framework was applied to the BioPAX version of two resources (PID,KEGG) of the Pathway Commons database and to the Atlas of Cancer Signalling Network (ACSN). As a case-study, it was used to characterize the causal signatures of markers of the epithelial-mesenchymal transition.

- *Functional signature for ADAMTS.* Hepatic Stellate Cells produce a wide variety of molecules involved in ECM remodeling, such as adamalysins [84]. However, the limitations of discovering new functions of these proteins stem from the experimental approaches that are difficult to implement due to their structure and biochemical features. In that context we developed an original framework combining the identification of small modules in conserved regions independent of known domains and the concepts of phylogenomics (association of conservation and phenotype gained concurrently during evolution). The resulting evolutionary model of motif signatures and protein-protein interaction signatures of the ADAMTS family is validated by data from literature and provides biologists with many new potential functional motifs [62], [63], [35].

- *Dynamic model of hepatic stellate cells.* To characterize the dynamics of HSC activation upon TGFB1 stimulation, we developed a model using Kappa, a site graph rewriting language and its static analyzer Kasa [56]. We previously demonstrated the advantages of Kappa language for modeling TGF-β signaling and extracellular matrix [85]. Unlike previous model based on a population of interacting proteins, we now develop an original Kappa model based on a population of cells interacting with TGF-β [30]. The model recapitulates the dynamics of activation of HSC towards myofibroblast states and the reversion processes. Current work aims to identify the regulators of the repair likely to promote the resolution of fibrosis at the expense of its progression.

5 Social and environmental responsibility

5.1 Footprint of research activities

Dyliss research activities have low environmental footprints. Most of our software solution run on off-the-shelf computers and are not computationally intensive. Indirectly, the analyses and predictions we

[13] cadbiom.genouest.org

make intend to reduce the need for long, costly technically or ethically difficult biological experiments.

5.2 Impact of research results

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

6 Highlights of the year

6.1 organization of the JOBIM conference

The Dyliss team has been heavily involved in the organization of the francophone bioinformatics conference JOBIM-2022 in Rennes (July 5th–8th): chair of conference committee (E. BECKER) and members of the organizing committee (C. BELLEANNÉE, N. BUTON, F. COSTE, O. DAMERON, O. DENNLER, J. GOT, C. JUIGNÉ, Y. LE CUNFF, B. RUIZ, N. THÉRET, K. THUILLIER, Y. TIRLET).

Dyliss also co-organized two of five mini-symposia during JOBIM-2022: *Management and integration of agronomic, phenotypic and environmental data* (O. DAMERON) and *Bioinformatics of metabolic pathways, from sequences to molecules* (J. GOT), with international invited speakers.

6.2 Reproducibility

We also note the emergence of a transversal theme on reproducibility, supported by several publications:

- Addressing barriers in comprehensiveness, accessibility, reusability, interoperability and reproducibility of computational models in systems biology [18] (A. Siegel)
- Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases [17] (M. Melkonian, C. Juigné, O. Dameron, E. Becker)
- Highlight on Semantic Web technologies are effective to remove redundancies from protein-protein interaction databases and define reproducible interactomes [29] (M. Melkonian, C. Juigné, O. Dameron, E.Becker)
- Improving reusability along the data life cycle: a Regulatory Circuits Case Study [16] (M. Louarn, X. Garnier, A. Siegel, O. Dameron)
- Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX [27] (C. Juigné, O. Dameron, F. Moreews, E. Becker)

6.3 Connecting the metabolic and regulatory scales

The methods developed in collaboration with Frei Berlin Univ, INRAE Toulouse and LABRI to learn regulatory rules controling metabolism were presented at the European Conference of Bioinformatics (ECCB) [21].

7 New software and platforms

7.1 New software

7.1.1 AskOmics

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

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

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

News of the Year: 2022: (1) Correction of various bugs, improvement of the user interface, and optimizations. (2) Modification of the description of relations and attributes are described. Currently blank nodes are used to avoid "phantom" relations and attributes. (3) Adding a management of ontologies and prefixes (possibilities of loading a part of NCBITAXON, for example going to search all descendants of a class). (4) Adding a "single tenant" mode. All graphs are public by default. Query performances have been significantly improved.

URL: https://askomics.org/

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

Contact: Olivier Dameron

Partners: Université de Rennes 1, CNRS, INRA

7.1.2 Metage2Metabo

Keywords: Metabolic networks, Microbiota, Metagenomics, Workflow

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

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

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

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

Publication: hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Arnaud Belcour, Anne Siegel

7.1.3 CADBIOM

Name: Computer Aided Design of Biological Models

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

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

News of the Year: We have significantly refactored Cadbiom package towards a framework that allows the identification of causal regulators in large-scale models, formalized in the BioPAX language and automatically interpreted as guarded transitions.

URL: http://cadbiom.genouest.org

Publications: inserm-00978313, hal-01242893, hal-01559249, hal-03693653

Contact: Anne Siegel

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

7.1.4 pax2graphml

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

Keyword: Bioinformatics

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

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

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

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

Publication: hal-03265223

Contact: François Moreews

Partner: INRAE

7.1.5 Protomata

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

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

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

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

The three main modules integrated in the Protomata-learner workflow are available as well as stand-alone programs: 1) paloma builds partial local multiple alignments, 2) protobuild infers automata from these alignements and 3) protomatch and protoalign scans, parses and aligns new sequences with learnt automata. The suite is completed by tools to handle or visualize data and can be used online by the biologists via a web interface on Genouest Platform.

News of the Year: Final release of paloma-2 (new and faster version of paloma in modern C++) introducing a new format for partial local alignments.

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

Contact: François Coste

Participant: François Coste

Partners: Université de Rennes 1, CNRS, Inria

7.1.6 PPsuite

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

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

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

News of the Year: A new Potts model inference method was developed, replacing the call to the external package CCMpredPy. This method is considerably faster, makes it possible to represent deeper and longer multiple sequence alignments, and is less sensitive to small sampling variations. To better handle insertion and deletions events, trimmed positions can now be represented by zero insertion penalties, enabling PPalign to return full alignment of the sequences, and the offset hyperparameter is now automatically determined from pseudocount rate and background amino acid frequencies.

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

Publications: hal-02402646, hal-02862213, hal-03264248, hal-03926272

Contact: François Coste

Participants: François Coste, Hugo Talibart, Mathilde Carpentier

7.1.7 Transformer Framework for Protein Characterization

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

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

Functional Description: Given examples of annotated sequences, this tool allows to train and analyse resulting models with respect to evaluation metrics (accuracy, correlation) plots. The process is fully automated and the whole operation can be done by modifying a JSON configuration file and providing a JSON data set. No code skills are thus required.

News of the Year: We added the ability to obtain attention maps during inference and combine these to get residue importance scores. We included in the framework eight additional interpretability methods that provide residue importance scores from the literature: Attention last layer, LIME, GradCam, LRP, Rollout, Gradients, InputXGrad, and Integrated gradient. And we implemented the computation of metrics enabling to evaluate the residue importance scores with respect to a ground truth target: Precision, Recall, Average Precision, Recall Gain, Precision gain, Maximum f-gain, Precision Recall Gain Area Under the Curve.

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

Contact: Nicolas Buton

Participants: Nicolas Buton, François Coste, Yann Le Cunff

7.1.8 Emapper2GBK

Keywords: Bioinformatics, Metabolic networks, Functional annotation

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

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

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

Publication: hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Arnaud Belcour, Anne Siegel

7.1.9 AuCoMe

Name: Automatic Comparison of Metabolisms

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

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

News of the Year: 2022: Transforming command merge into spontaneous. Adding two filtering options: –filtering –union and –filtering –intersection in the orthology step. Improving Pypi and readme file.

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

Contact: Anne Siegel

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

7.1.10 mpwt

Keywords: Metabolic networks, Multi-processor

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

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

Publication: hal-02395024

Contact: Anne Siegel

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

8 New results

8.1 Scalable methods to query data heterogeneity

Participants: Emmanuelle Becker, Cécile Beust, Olivier Dameron, Xavier Garnier, Camille Juigné, Marine Louarn, Marc Melkonian, Francois Moreews, Anne Siegel, Nathalie Théret, Yael Tirlet.

Data engineering: designing an integrative data model for heterogeneous data [*C. Beust, O. Dameron, N. Théret, Y. Tirlet*] [31, 25, 48]

- A book chapter presents the main challenges of integrating and reasoning with life science data, and surveys how Semantic Web technologies are a relevant framework for addressing these issues. [31]

- During Cécile Beust's internship, we converted the Comparative Toxicogenomics Database (CTD) into the BioPAX format. We expect to use the result as an input for CadBiom in order to build large-scale biological dynamic networks based on guarded transitions. The generated models could be analyzed thanks to reachability queries in order to identify environmental exposures causal signatures associated to the occurrence of chronic liver diseases. [42]

- Many studies focus on phenotype observation of species of interest, with additional data on experimental conditions and on metagenomic for biodiversity. We designed a data schema that (1) avoids the unnecessary duplication of data engineering for each study, (2) provides a common repository of queries and (3) in the long run supports combining data from multiple studies. We deployed this data model using AskOmics. We validated our approach by integrating the data from a previous article and repoducing the analysis. This work will serve as a foundation for our future contribution to the DeepImpact project. [25, 48]

Improving reusability along the data life cycle: a Regulatory Circuits Case Study [*O. Dameron, X. Garnier, M. Louarn, A. Siegel*] [16]

- Many life science studies' data are provided using specific and non-standard formats. This hampers the capacity to reuse the studies data in other pipelines, the capacity to reuse the pipelines results in other studies, and the capacity to enrich the data with additional information. We designed a modular RDF representation of the Regulatory Circuits data, the sample-specific and the tissue-specific networks, and the corresponding metadata. The result is available at zenodo[14]. It supports biologically-relevant SPARQL queries. It allows an easy and fast querying of the resources related to the initial Regulatory Circuits datasets and facilitates its reuse in other studies.

Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases [*E. Becker, O. Dameron, C. Juigné, M. Melkonian*] [17, 29, 33]

- Information on protein-protein interactions is collected in numerous primary databases. Several meta-databases aggregate primary databases to provide more exhaustive datasets. Redundancy occurs in meta-databases when some publications reporting protein-protein interactions have been annotated with different precision levels by multiple primary databases. We proposed a precise definition of explicit and implicit redundancy, and showed that both can be easily detected using Semantic Web technologies. We applied this process to a dataset from the APID meta-database and showed that while explicit redundancies were detected by the APID aggregation process, about 15% of APID entries are implicitly redundant and should not be taken into account when presenting confidence-related metrics. Finally, we built a "reproducible interactome" with interactions that have been reproduced by multiple methods or publications. The size of the reproducible interactome was drastically impacted by removing redundancies for both yeast (-59%) and human (-56%), and we showed that this is largely due to implicit redundancies.

- This work was also a part of the habilitation thesis "From homogeneous data to heterogeneous data in systems biology" defended by Emmanuelle Becker [33]

Extracting robust information from BioPAX databases. [*E. Becker, O. Dameron, C. Juigné, F. Moreews, A. Siegel, N. Théret*] [26, 27, 33, 22]

- Molecular complexes play a major role in the regulation of biological pathways. The Biological Pathway Exchange format (`BioPAX`) facilitates the integration of data sources describing interactions some of which involving complexes. The `BioPAX` specification explicitly prevents complexes to have any component that is another complex. However, we observed that the well-curated Reactome pathway database contains such recursive complexes of complexes. We proposed reproductible and semantically-rich SPARQL queries for identifying and fixing invalid complexes in `BioPAX` databases, and evaluate the consequences of fixing these non-conformities in the Reactome database [27]. Overall, this method improved the conformity and the automated analysis of the graph by repairing the topology of the complexes in the graph. This will allow to apply further reasoning methods on better consistent data.

[14] `zenodo:4889146`

- A large quantity of experimental data can be generated on tissues and cells by using complementary high-throughput techniques like transcriptomics, proteomics and metabolomics, as well as by target analyses for specific molecules. This results in a large amount of multimodal data. Each modality can be statistically analyzed to produce lists of differentially-expressed molecules between experimental conditions. We hypothesized that considering the different levels of omics as a whole will help understand biological systems, and introduced a methology to map results of high-throughput transcriptomic and high-throughput metabolomic data on a graph representing metabolism (interactions and their regulation [26]. This graph highlights the links between small molecules and proteins, and thus might be the appropriate system to integrate both metabolomic and transcriptomic data. This work opens new perspectives to integrate simultaneously proteomic/transcriptomic and metabolomic data, and to find networks between these molecules or potential common upstream regulators.

- This work was also a part of the habilitation thesis "From homogeneous data to heterogeneous data in systems biology" defended by Emmanuelle Becker [33]

- Although the BioPAX standard has been widely adopted by the community to describe biological pathways, no computational method is able of studying the dynamics of the networks described in the BioPAX large-scale resources. To solve this issue, our Cadbiom framework was designed to automatically transcribe the biological systems knowledge of large-scale BioPAX networks into discrete models. The framework then identifies the trajectories that explain a biological phenotype (e.g., all the biomolecules that are activated to induce the expression of a gene) [22].

Addressing barriers in comprehensiveness, accessibility, reusability, interoperability and reproducibility of computational models in systems biology. [*A. Siegel*] [18]

- Computational models are often employed in systems biology to study the dynamic behaviours of complex systems. With the rise in the number of computational models, finding ways to improve the reusability of these models and their ability to reproduce virtual experiments becomes critical. Correct and effective model annotation in community-supported and standardised formats is necessary for this improvement. Here, we present recent efforts toward a common framework for annotated, accessible, reproducible and interoperable computational models in biology, and discuss key challenges of the field.

8.2 Metabolism: from protein sequences to systems ecology

Participants: Arnaud Belcour, Samuel Blanquart, Nicolas Buton, François Coste, Jeanne Got, Pauline Hamon-Giraud, Yann Le Cunff, Victor Mataigne, Anne Siegel, Nathalie Théret, Yael Tirlet.

Modeling proteins with crossing dependencies [*F. Coste*] [47]

- In collaboration with H. Talibart and M. Carpentier (ISYEB, Muséum national d'Histoire naturelle), we proposed a new Potts model inference method that is considerably faster, enabling to represent and align deeper and larger alignments, and that can use pseudocounts to improve the robustness of the inference with respect to sampling variations. This method has been implemented in the PP suite [47].

Deep attention networks for enzyme class predictions [*N. Buton, F. Coste, Y. Le Cunff*] [43]

- We studied the interest of Transformer deep neural networks for the functional annotation of sequences by focusing on the prediction of enzymatic classes. Our EnzBert transformer models, trained to predict enzyme commission (EC) numbers by specialization of a protein language model, were able to significantly outperform state-of-the-art tools for monofunctional enzyme class prediction based on sequences only. We also showed that the attention of Transformers provides

an interesting built-in mechanism for the interpretabitlity of these predictions by proposing a simple aggregation of the attention maps which was on par with, or better than, other classical interpretability methods on predicting the enzymatic sites of enzymes [43].

Large-scale eukaryotic metabolic network and design of microbial communities [*A. Siegel, A. Belcour, S. Blanquart, J. Got, N. Théret, V. Mataigne, Y. Tirlet, P. Hamon-Giraud*] [34, 13, 23, 20, 32].

- *Combining knowledge-based and sequence comparison approaches to elucidate metabolic functions* [34] The taxonomic diversity within an environmental sample is most often identified through targeted gene marker sequencing, or amplicon. We designed a new method estimating the metabolic capacities of a wild organism based on the estimated taxonomy of its sequenced amplicon sequence. The method consists in selecting taxonomicaly close annotated genomes in UniProt, then it estimates clusters of shared enzymes to identify the core proteome of the taxon. The core-proteome can be considered as a proxy of the wild organim metabolic capacities. Unlike other approaches in this field, our method considers taxonomic assignments as inputs and not exclusively 16S rRNA amplicons, and it provides as output a metabolic network instead of a metabolic profiling. The method, implemented in the Esmecata tool, and its application to the biogaz reactor are described in Arnaud Belcour PhD thesis [34].

- *Insights into the potential for mutualistic and harmful host-microbe interactions affecting brown alga freshwater acclimation* [13] Microbes can modify their hosts' stress tolerance, thus potentially enhancing their ecological range. An example of such interactions is *Ectocarpus subulatus*, one of the few freshwater-tolerant brown algae. This tolerance is partially due to its (un)cultivated microbiome. The biological station of Roscoff investigated this phenomenon by modifying the microbiome of laboratory-grown *E. subulatus* using mild antibiotic treatments, which affected its ability to grow in low salinity. Low salinity acclimation of these algal-bacterial associations was then compared, including a study at the metabolic scale using the tool designed in the Dyliss team, and reviewed in [32]: gene expression of the host and metabolite profiles were affected almost exclusively in the freshwater-intolerant algal-bacterial communities, and vitamin K synthesis is one possible bacterial service missing specifically in freshwater-intolerant cultures in low salinity. Together, these results provide two promising hypotheses to be examined by future targeted experiments.

- *Microbial genomics: from cells to genes (and back to cells)* [20] The rumen harbours countless various microorganisms that have established multiplicity of relationships to efficiently digest complex nutrients, essentials for the host's health, growth and performances. Recent studies using omics-based techniques have revealed that changes in rumen microbiota are associated with changes in ruminants' production and health parameters. This review advocate for the benefits of switching from traditional rumen microbes studies using anaerobic culture-based techniques to molecular techniques applied to microbial cultures. The paper provides a comprehensive review of current advances in molecular methods to identify novel rumen microbes and discuss how culturing and mathematics could enhance our understanding of rumen microbiology.

- *Input contributions on metabolic outputs : application to human diets* [23] The public availability of human microbiome datasets makes it possible to apply diets to these human microbiomes metabolim to model the behavior of organisms. We automated an approach (nAIO) that allows, for each input nutrient in the network, to determine the percentages that are distributed in the different outputs when the organism is forced to evolve in a given diet. The nAIO is computed thanks to the inversion of a large-scale matrix and is combined with linear optimization problems. We applied this method to all known bacterial networks from studies of the gut microbiota and stored in the Virtual Metabolic Human database. The calculation of nAIOs shows that computation times do not depend on the size of the network but rather on the selected diet. The nAIO calculation also shows that for some bacteria the nAIOs are independent of diet. For these bacteria the nAIOs can be used to make predictions that result in a linear relationship between the inputs of the system and its outputs.

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

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

Learning Boolean controls in regulated metabolic networks: a case-study [*A. Siegel, K. Thuillier*] [21]

- Many techniques have been developed to infer Boolean regulations from a prior knowledge network and experimental data. Existing methods are able to reverse-engineer Boolean regulations for transcriptional and signaling networks, but they fail to infer regulations that control metabolic networks. We present a novel approach to infer Boolean rules for metabolic regulation from time series data and a prior knowledge network. Our method is based on a combination of answer set programming and linear programming and generates candidate Boolean regulations that can reproduce the given data when coupled to the metabolic network. We evaluated our approach on a core regulated metabolic network and show how the quality of the predictions depends on the available kinetic, fluxomics or transcriptomics time series data.

Discrete modeling for integration and analysis of large-scale signaling networks [*A. Siegel, N. Théret*] [22]

- The computation of sets of biological entities implicated in phenotypes is hampered by the complex nature of controllers acting in competitive or cooperative combinations. The identification of controllers relies on computational methods for dynamical systems, which require the biological information about the interactions to be translated into a formal language. We used the `biopax2cadbiom` method to create Cadbiom models from three biological pathway databases (KEGG, PID and ACSN). The cadbiom framework then identifies the trajectories that explain a biological phenotype (e.g., all the biomolecules that are activated to induce the expression of a gene). The comparative analysis of these models highlighted the diversity of molecules in sets of biological entities that can explain a same phenotype. The application of our framework to the search of biomolecules regulating the epithelial-mesenchymal transition not only confirmed known pathways in the control of epithelial or mesenchymal cell markers but also highlighted new pathways for transient states.

Functional signature for ADAMTS [*C. Belleannée, S. Blanquart, F. Coste, O. Dennler, N. Théret*] [44, 35].

- Hepatic Stellate Cells produce a wide variety of molecules involved in ECM remodeling, such as adamalysins [84]. However, the limitations of discovering new functions of these proteins stem from the experimental approaches that are difficult to implement due to their structure and biochemical features. In that context, we developed an original framework combining the identification of small modules in conserved regions independent of known domains and the concepts of phylogenomics (association of conservation and phenotype gained concurrently during evolution). We estimated the phylogenetic history of ADAMTS and ADAMTS like proteins in nine bilateria species including human, suggesting the emergence of the ADAMTSL and papilin within the ADAMTS. A dataset of 447 protein-protein interactions (PPI) with the 26 ADAMTS-TSL human paralogs was constructed and we estimated ancestral scenario for PPI appearances along our bilateria tree. We found 45 ancestors displaying a co-appearance of conserved module signatures and PPI. We identified convergent appearances of PPI with COMP and CCN2 and we showed that distinct signatures of the ADAMTS7, ADAMTS3 and ADAMTS4 ancestors could be involved in those interactions. We finally obtained a signature discontinuous along the primary sequence but folding in a contiguous three dimensional region in the hyalectanase sub-group of ADAMTS and putatively involved in the ACAN and VCAN interactions. The resulting evolutionary model of motif signatures and protein-protein interaction signatures of the ADAMTS family is validated by data from literature and provides biologists with many new potential functional motifs freely available on ITOL.Olivier Dennler defended his PhD thesis [35] and an article is under consideration by an international journal.

Creation of predictive functional signaling networks [*M. Bougueon, N. Théret*] [30].

- *The rule-based model approach. A Kappa model for hepatic stellate cells activation by TGFB1* [30] Kappa is a site graph rewriting language. It offers a rule-centric approach, inspired from chemistry, where interaction rules locally modify the state of a system that is defined as a graph of components, connected or not. In this case study, the components will be occurrences of hepatic stellate cells in different states, and occurrences of the protein TGFB1. The protein TGFB1 induces different behaviors of hepatic stellate cells thereby contributing either to tissue repair or to fibrosis. Better understanding the overall behavior of the mechanisms that are involved in these processes is a key issue to identify markers and therapeutic targets likely to promote the resolution of fibrosis at the expense of its progression.

Characterizing gene structure with grammatical languages and conservation information [*C. Belleannée, S. Blanquart, O. Dameron, N. Guillaudeux*] [12]

- Based on syntactic models and graph formalisms, we compared splicing structures of 2167 triplets of orthologous genes shared in human, mouse and dog. This resulted in the prediction of 6861 new coding transcripts (*i.e.* putative proteins) on these species, mainly for dog, an emergent model species. Every predicted transcript shares an identical exonic structure with a coding transcript already known in another species, hence defining them as orthologs. Additionnaly, we identified a set 253 gene triplets with strictly conserved exonic structures in human, mouse and dog, and so expressing the same proteome (*i.e.* the same isoform coding transcripts). These genes express a total of 879 groups of orthologous isoforms, such that in each group, the same splicing structure is shared in each three species gene. Although these genes express a same proteome, we showed that the expressed transcriptomes may be different, due to the gene's propensity to express distinct alternatively transcribed mRNAs encoding the same protein [12].

Signatures for fronto-temporal degeneration and amyotrophic lateral sclerosis [E. Becker, V. Kmetzsch] [15, 14, 24, 36, 33]

- In the context of our participation in the IPL NeuroMarker project, a joint study with Institut du Cerveau (Inseerm/CNRS/Sorbonne Université) at the Pitié-Salpêtrière hospital and the Aramis team (Inria Paris) evidenced a signature of four microRNAs in n presymptomatic and symptomatic subjects with frontotemporal dementia and amyotrophic lateral sclerosis associated with a C9orf72 mutation [73]. To critically assess the discriminative power of this signature and compare it with other available signatures, this study was followed with a validation study that highlighted the discriminative power of the different signatures with an independant cohort [15]. This large scale reproducibility analysis of previously identified microRNA signatures for fronto-temporal degeneration and amyotrophic lateral sclerosis was mostly confirmed the signature identified in [73] as discriminative for patients and pre-symptomatic carriers of the C9orf72 mutation.

- In recent years, many approaches have been developed for modeling disease progression from data, most of these approaches requiring longitudinal data. Indeed, among the proposed models, only the event-based models (EBMs) allow inferring a disease progression score from cross-sectional data. However, EBMs were applied to a relatively small number of variables (typically 10-50) and it is not known whether they perform well in higher dimensions. In the context of the IPL NeuroMarker, we proposed a new model using cross-sectional multimodal data based on variational autoencoders in 3 steps: (1) estimation of the latent space, (2) definition in the latent space of a curve defining the progression of the disease; and finally (3) estimation of the progression score of an individual by orthogonal projection of its coordinates in the latent space onto the main trajectory curve. When applied to frontotemporal degeneration and amyotrophic lateral sclerosis, the proposed method was more efficient than EBMs [14, 24, 28].

- Virgilio Kmetzsch also defended his PhD thesis "Multimodal analysis of neuroimaging and transcriptomic data in genetic frontotemporal dementia". [36], and this work was also a part of the habilitation thesis "From homogeneous data to heterogeneous data in systems biology" defended by Emmanuelle Becker [33]

Signatures of mutants of the enzyme EXOSC10/Rrp6 [E. Becker] [19]

- The conserved 3'-5' exoribonuclease EXOSC10/Rrp6 is required for gametogenesis, brain development, erythropoiesis and blood cell enhancer function. The human ortholog is essential for mitosis in cultured cancer cells. Little is known, however, about the role of Exosc10 during embryo development and organogenesis. The transcriptional landscape of EXOSC10 mutants was investigated to explain its essentiality for the eight-cell embryo/morula transition [19].

Signature of Crohn's disease symptoms from microbiota profiles[Y. Le Cunff] [11]

- Standard approaches to describe patients' microbiota in Crohn's disease (CD) consists in comparing those with control individuals' microbiota. In this work, we decided to rather focus on distinguishing subgroups of microbiota profiles within a novel CD cohort studied in Rennes' C.H.U. We used unsupervised clustering techniques to highlight the existence of three microbiota subprofiles, each linked with a different symptoms' severity. Moreover, we also showed that these groups are largely stable over time. Finally, using differential abundance analysis, we managed to point out key species which could act as signatures for CD evolution over time. [11]

9 Partnerships and cooperations

9.1 International initiatives

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

SymBioDiversity

Title: Symbolic and numerical mining and exploration of functional biodiversity

Duration: 2020 – 2024

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

Partners:

- Universidad de Chile (Chili)

Inria contact: Anne Siegel

Summary: The project aims at developing methods combining data-mining, reasoning and mathematical modeling to efficiently analyze massive data about microbial biodiversity in extreme environment and identify families of species which characterize environmental niches. The partnership combines Inria Team Dyliss (systems biology, reasoning), Pléiade (systems biology, biodiversity), the chilean Center of Mathematical Modeling (modeling of ecosystems), Inria Chile (data mining, transfer) and chilean biologist partners experts in biodiversity (universidad catholica). As the team started in 2020 during the pandemic crisis, Inria decided to extend it for three additional years.

9.2 International research visitors

9.2.1 Visits of international scientists

Other international visits to the team

Oumarou Abdou-Arbi

Status: Researcher

Institution of origin: Université Dan Dicko Dankoulodo Maradi

Country: Niger

Dates: Aug. to Sept. 2022

Context of the visit: This visit resulted in a communication [23]

Mobility program/type of mobility: Research stay

Alejandro Maass

Status: Researcher

Institution of origin: Universidad de Chile

Country: Chile

Dates: Oct. 2022

Context of the visit: SymBioDiversity associated Inria team

Mobility program/type of mobility: Research stay

9.2.2 Other european programs/initiatives

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

Participants: Moana Aulagner , Samuel Blanquart , Anne Siegel .

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

9.3 National initiatives

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

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

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

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

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

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

IDEALG (ANR/PIA-Biotechnology and Bioresource)

Participants: Arnaud Belcour , François Coste , Jeanne Got , Anne Siegel .

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)[15]. 2012–2021. Total grant: 11M€. Dyliss grant: 534k€.

PhenomiR

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

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

9.3.1 Programs funded by Inria

IPL Neuromarkers

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

[15] idealg.u-bretagneloire.fr/

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

9.4 Regional initiatives

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

Participants: François Coste.

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

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- ICGI 2023 (International Conference on Grammatical Inference), Rabat, Morocco [F. Coste]

Member of organizing committees

- Jobim 2022 (Journées Ouvertes Biologie Informatique Mathématiques, 500 participants) [C. Belleannée, N. Buton, F. Coste, O. Dameron, O. Dennler, J. Got, C. Juigne, Y. Le Cunff, B. Ruiz, N. Theret, K. Thuillier, Y. Tirlet]

Member of competition scientific committees

- TASSIL (TrAnsformers And recurrent neural networkS diStiLlation Challenge) [F. Coste]

10.1.2 Scientific events: selection

Chair of conference program committees

- Jobim 2022 (Journées Ouvertes Biologie Informatique Mathématiques), Rennes, France [E. Becker]

Member of the conference program committees

- ISMB-2022 (International Symposium on Molecular Biology) [A. Siegel]
- ICML'2022 workshop WBC [A. Siegel]
- OnUCAI-KR2022 (Ontology Uses and Contribution to Artificial Intelligence) [O. Dameron]
- Jobim 2022 (Journées Ouvertes Biologie Informatique Mathématiques), Rennes, France [O. Dameron, A. Siegel]
- Journée Santé et IA, satellite of PFIA Plate-Forme Intelligence Artificielle [O. Dameron]

Reviewer

- Jobim 2022 [O. Dameron]
- IA et santé 2022 [O. Dameron]

10.1.3 Journal

Member of the editorial boards

Reviewer - reviewing activities

- Bioinformatics [F. Coste]
- Briefings in Bioinformatics [O. Dameron, x4]

10.1.4 Invited talks

- Workshops "Le RNA-seq, de la paillasse à l'analyse in-silico", PLBS UAR, Lille [S. Blanquart]
- Journée de lancement du PEPR Atlasea " Des génomes marins à la synthèse in silico et in vivo de molécules" [A. Siegel]

10.1.5 Scientific expertise

Local responsibilities

- Organisation of the bioinformatics teams (Dyliss, GenOuest and GenScale as well as members of other bioinformatics teams in Rennes) weekly seminars [S. Blanquart]
- Chargé de mission "Numérique et Environnement" for Inria centre at Rennes University [S. Blanquart]
- Chargé de mission "Biologie et Santé Numériques" for Inria centre at Rennes University [F. Coste]
- Scientific Advisory Board of the GenOuest platform [O. Dameron]
- Responsability of the IRISA laboratory "Health-biology" cross-cutting axis [Y. Le Cunff]
- Scientific Advisory Board of Biogenouest [N. Théret]
- Delegate to research integrity at the University of Rennes 1 [N. Théret]

10.1.6 Research administration

Institutional boards for the recruitment and evaluation of researchers

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

Scientific councils

- Scientific referent (for CNRS) of the PEPR exploratoire Molecularxiv [A. Siegel]
- Comité de pilotage of the Mission for Interdisciplinarity (MITI) at CNRS
- Scientific council of the PPR Autonomy [A. Siegel]
- Member of the coordination comitee of the ModCov19 research group [A. Siegel]
- Scientific council of the MathNum department of Inrae [A. Siegel]

National responsibilities

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

Local responsibilities

- Member of the Inria Rennes center council [J. Got]
- Member of the Biology department council [Y. Le Cunff]
- CUMI (Commission des utilisateurs des moyens informatiques) of Inria Rennes [F. Coste]
- Social committee of Univ. Rennes 1 [C. Belleannée]
- Emergency aid commission of Univ. Rennes 1 and Rennes 2 [C. Belleannée]

10.2 Teaching - Supervision - Juries

10.2.1 Teaching tracks responsibilities

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

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]
- "Imperative programming", Licence 1 informatique, Univ. Rennes 1 [O. Dameron]
- "Complément informatique 1", Licence 1 informatique, Univ. Rennes 1 [O. Dameron]
- "Atelier bioinformatique", Licence 2 informatique, Univ. Rennes 1 [O. Dameron]
- "Semantic Web and bio-ontologies", Master 2 in bioinformatics, Univ. Rennes 1 [O. Dameron]
- "Internship", Master 2 in bioinformatics, Univ. Rennes 1 [O. Dameron]
- "Integrative and Systems biology", Master 2 in bioinformatics, Univ. Rennes 1 [A. Siegel]

- "Micro-environnement Cellulaire normal et pathologique", Master 2 Biologie cellulaire et Moléculaire, Univ. Rennes 1 [N. Théret]
- "Machine Learning", Master 1 in Bioinformatics [Y. le Cunff]
- "Modeling dynamic systems", Licence 2, Biology [Y. Le Cunff]
- "Simulating Biological Systems", Master 2 in Bioinformatics [Y. Le Cunff]
- "Simulation and biology interfaces", Master 1 in Biology [Y. Le Cunff]
- "Applied interdisciplinarity", Master 2 in Biology [Y. Le Cunff]
- "Introduction to Machine Learning for biology", Ph.D. Program [Y. Le Cunff]

10.2.3 Teaching

Relecture Anaïs : on est encore en train de finaliser cette section

- Master : E. Becker, "Introduction to computational ecology", 34h, Master 2 in Ecology, Univ. Rennes 1, France
- Master : E. Becker, "Method", 15h, Master 2 in Computer Sciences, Univ. Rennes 1, France
- Master : E. Becker, "Insertion Professionnelle et tables rondes", 6h, Master 1 and Master 2 in Bioinformatics, Univ. Rennes 1, France
- Master : E. Becker, "Systems Biology : biological netorks", 27h, Master 2 in Bioinformatics, Univ. Rennes 1, France
- Master : E. Becker, "Introduction to Bioinformatics", 3h, Master MEEF Biology, Univ. Rennes 1, France.
- Licence: C. Belleannée, Langages formels, 20h, L3 informatique, Univ. Rennes 1, France.
- Licence: C. Belleannée, Projet professionnel et communication, 16h, L1 informatique, Univ. Rennes 1, France.
- Licence: C. Belleannée, Enseignant référent, 20h, L1 informatique, Univ. Rennes 1, France.
- Licence: C. Belleannée, Spécialité informatique : Functional and immutable programming , 42h, L1 informatique, Univ. Rennes 1, France
- Master: C. Belleannée, Algorithmique du texte et bioinformatique, 10h, M1 informatique, Univ. Rennes 1, France
- Master: C. Belleannée, Programmation logique et contraintes, 32h, M1 informatique, Univ. Rennes 1, France
- Licence: C. Belleannée, Outils formels pour l'informatique, 32h, L2 informatique, Univ. Rennes 1, France
- Master: F. Coste, Supervised machine learning, 10h, M2 Science Informatique, Univ. Rennes, France
- Licence: O. Dameron, "Programmation 1", 40h, Licence 1 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Complément informatique", 24h, Licence 1 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Atelier bioinformatique", 24h, Licence 2 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Databases", 24h, Licence 2 informatique, Univ. Rennes 1, France

- Licence: O. Dameron, "Programmation", 54h, Licence 3 miage, Univ. Rennes 1, France
- Master: O. Dameron, "Semantic Web", 20h, Master 1 miage, Univ. Rennes 1, France
- Master: O. Dameron, "Veille technologique", 2h, Master 2 miage, Univ. Rennes 1, France
- Master: O. Dameron, 2h, "Internship", Master 1 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, 20h, "Semantic Web and bio-ontologies", Master 2 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, 18h, "Internship", Master 2 in bioinformatics, Univ. Rennes 1, France
- Master: A. Siegel, Integrative and Systems biology, Master 2 in bioinformatics, Univ. Rennes 1.
- Licence : Y. Le Cunff "Modélisation des phénomènes du vivant", 30h, L2 Biologie, Univ. Rennes 1, France
- Master: Y. Le Cunff, "Apprentissage statistique", 110h, Master 1 in Bioinfortmatics Univ. Rennes 1, France
- Master: Y. Le Cunff, "Biologie aux interfaces", 25h, Master 1 in Biology, Univ. Rennes 1, France
- Master: Y. Le Cunff,"Simulating dynamic systems in biology", Master 2 in bioinformatics, 20h, Univ. Rennes 1, France
- Master: Y. Le Cunff, "Applied Interdisciplinarity", 20h, Master 2 in biology, Univ. Rennes 1, France
- PhD program: Y. Le Cunff,"Introduction to Machine Learning", 20h, FdV PhD Program, Sorbonne Paris Université, Paris, France

10.2.4 Supervision

- PhD: Arnaud Belcour, *Combining knowledge-based and sequence comparison approaches to elucidate metabolic functions, from pathways to communities*, defended Oct. 21st 2022, supervised by A. Siegel and S. Blanquart. [34]
- PhD: Olivier Dennler, Modular functional characterization of ADAMTL and ADAMTSL protein families, defended Dec. 19th 2022, supervised by N. Theret, F. Coste, S. Blanquart and C. Belleannée. [35]
- PhD: Virgilio Kmetzsch, *Multimodal analysis of neuroimaging and transcriptomic data in genetic frontotemporal dementia*, defended 26th 2022, co-supervised by E. Becker and O. Colliot (Inria ARAMIS). [36]
- PhD in progress: Matthieu Bouguéon, *Modélisation prédictive pour le ciblage thérapeutique du TGF-beta dans les pathologies chroniques hépatiques*, started in Oct. 2020, supervised by N. Théret and A. Siegel
- PhD in progress: Nicolas Buton, *Deep learning for proteins functional annotation : novel architectures and interpretability methods*, started in Oct. 2020, supervised by F. Coste, Y. Le Cunff and O. Dameron.
- PhD in progress: Camille Juigné, *Analyse des données biologiques hétérogènes par exploitation de graphes multicouches pour comprendre et prédire les variations d'efficacité alimentaire chez le porc*,started in Dec. 2020, supervised by E. Becker and F. Gondret (INRAE Pegase)
- PhD in progress: Baptiste Ruiz *Algorithmes d'apprentissage automatique appliqués au microbiote : Intégration de connaissances a priori pour de meilleures prédictions de phénotype*, started in Oct. 2021, supervised by Y. Le Cunff and A. Siegel

- PhD in progress: Kerian Thuillier *Inférence de règles booléennes contrôlant des modèles hybrides de systèmes biologiques multi-échelles*, started in Oct. 2021, supervised by A. Siegel and L. Paulevé (LABRI)

- PhD interrupted in 2022: Marc Melkonian *Intégration de données et de connaissances pour l'analyse fine de l'interactome*, started in Dec. 2021, supervised by E. Becker and G. Rabut (IGDR) [17, 29]

- L3 ENS internship: Thibaut Antoine, Pondération d'automates obtenus par alignements partiels de séquences protéiques, May-July 2022, supervised by F. Coste [40]

- M1 internship: Moana Aulagner, *Draft reconstructions of metabolic networks from Nanopore and shotgun metagenomic data*, Apr.– Jul. 2022, co-supervised by A. Siegel; J. Got and E. Roux (NuMeCan) [41]

- M1 internship: Cécile Beust, *Search for exposomic causality of liver fibrosis using network analysis*, Apr.– Jul. 2022, co-supervised by N. Théret and O. Dameron [42]

- M2 internship: Pauline Hamon-Giraud, *Comparative analysis of genome scale metabolic networks in brown algae*, Jan.–Jul. 2022, co-supervised by A. Siegel, J. Got and G. Markov (Station Biologique de Roscoff) [45]

- M2 internship: Yael Tirlet, *Intégration de données agricoles et environnementales à l'aide des technologies du Web sémantique*, Jan.–Jul. 2022, co-supervised by M. Boudet and O. Dameron [48, 25]

10.2.5 Doctoral advisory committee (CSID)

- Manon Lesage, Université de Rennes [E. Becker]

- Guillaume Doré, Université de Rennes [E. Becker]

- L. Cornanguer, Université de Rennes 1 [F. Coste]

- Yvon Awuklu, Université de Bordeaux [O. Dameron]

- Marine Djaffardjy, Université Paris-Saclay [O. Dameron]

- Tiphaine Casy, Université de Rennes 1 [O. Dameron]

- Andreas Checcoli, Institut Curie [A. Siegel]

- Maxime Lecompte, Univ. Bordeaux [A. Siegel]

- Olivier Quenez, Univ. Caen [A. Siegel]

- Indusha Kughatas, Université de Rennes 1 [Y. Le Cunff]

- Camille Kergal, Université de Rennes 1 [Y. Le Cunff]

10.2.6 Juries

Referee of PhD thesis:

- M. Delmas, Univ. Toulouse [O. Dameron]

- J. Grignard, Univ Paris-Saclay [O. Dameron]

Member of PhD thesis juries:

- S. Pankaew, Université Aix-Marseille [E. Becker]
- S. le Bars, Univ Nantes [A. Siegel]
- V. Kmetzsch, Université Paris-Sorbonne [E. Becker, Y. Le Cunff]
- O. Dennler, Université de Rennes 1 [C. Belleannée, S. Blanquart, F. Coste, N. Theret]
- A. Belcour, Université de Rennes 1 [S. Blanquart, O. Dameron, A. Siegel]

Member of habilitation thesis juries:

- E. Becker, Université de Rennes 1 [O. Dameron, A. Siegel]
- A. Goezler, Univ Paris Saclay [A. Siegel]

10.3 Popularization

10.3.1 Articles and contents

Popularizing sciences at the national level Our team was involved in the cosupervision of a comic book gathering the portraits of 12 female computer scientists Les décodeuses du numérique. The book was sent to all French high schools and is freely available online[16]: more than 30,000 views and 5,000 downloads since September 2021. This led to several intervention in round-tables, large-audience conferences and media: Radio France International, Emission “autour de la question” - ONU info, Salon educatice-educatech - Round-table at the ESIR Engineer school, Rennes, “parité et les métiers du numérique au féminin”, 2021 - Women in AI conference, 2022 - Round- table Femmes dans l’informatique, Sorbonne Université - Semaine de l’égalité, Académie de Toulouse - Journée citoyenneté numérique, Academy of Créteil - Les rencontres de l’AIEF.

10.3.2 Interventions

- "Talkin’ Bout A ML Revolution", an introduction to Artificial Intelligence and Machine Learning at "Journée de l’axe Bio-informatique du réseau Biogenouest: L’intelligence Artificielle au service des sciences de la vie" [F. Coste]
- "Recherche de poste pour les jeunes bioinfos", panel JEBIF (French association of the young bioinformaticians) [O. Dameron]
- "Avenir de la bioinformatique", panel JEBIF (French association of the young bioinformaticians) [O. Dameron]

10.3.3 Contributions to open source projects

A. Belcour filled several issues and provided bugfixes to open source projects in connection with his research activity at DYLISS.

11 Scientific production

11.1 Major publications

[1] M. Aite, M. Chevallier, C. Frioux, C. Trottier, J. Got, M.-P. Cortés, S. N. Mendoza, G. Carrier, O. Dameron, N. Guillaudeux, M. Latorre, N. Loira, G. V. Markov, A. Maass and A. Siegel. ‘Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models’. In: *PLoS Computational Biology* 14.5 (May 2018). e1006146. DOI: 10.1371/journal.pcbi.1006146. URL: https://hal-univ-rennes1.archives-ouvertes.fr/hal-01807842.

[16] www.ins2i.cnrs.fr/fr/les-decodeuses-du-numerique

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

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

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

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

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

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

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

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

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

11.2 Publications of the year

International journals

[11] S. Buffet-Bataillon, G. Bouguen, F. Fleury, V. Cattoir and Y. Le Cunff. 'Gut microbiota analysis for prediction of clinical relapse in Crohn's disease'. In: *Scientific Reports* 12.1 (Dec. 2022), p. 19929. DOI: 10.1038/s41598-022-23757-x. URL: https://hal.science/hal-03868943.

[12] N. Guillaudeux, C. Belleannée and S. Blanquart. 'Identifying genes with conserved splicing structure and orthologous isoforms in human, mouse and dog'. In: *BMC Genomics* 23.1 (2022), pp. 1–14. DOI: 10.1186/s12864-022-08429-4. URL: https://hal.archives-ouvertes.fr/hal-03616626.

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

[14] V. Kmetzsch, E. Becker, D. Saracino, D. Rinaldi, A. Camuzat, I. Le Ber and O. Colliot. 'Disease progression score estimation from multimodal imaging and microRNA data using supervised variational autoencoders'. In: *IEEE Journal of Biomedical and Health Informatics* (2023), pp. 1–12. DOI: 10.1109/JBHI.2022.3208517. URL: https://hal.archives-ouvertes.fr/hal-03789357.

[15] V. Kmetzsch, M. Latouche, D. Saracino, D. Rinaldi, A. Camuzat, T. Gareau, I. Le Ber, O. Colliot and E. Becker. 'MicroRNA signatures in genetic frontotemporal dementia and amyotrophic lateral sclerosis'. In: *Annals of Clinical and Translational Neurology* (2022), pp. 1–14. DOI: 10.1002/acn3.51674. URL: https://hal.archives-ouvertes.fr/hal-03826747.

[16] ***Best Paper***
M. Louarn, F. Chatonnet, X. Garnier, T. Fest, A. Siegel, C. Faron and O. Dameron. 'Improving reusability along the data life cycle: a Regulatory Circuits Case Study'. In: *Journal of Biomedical Semantics* 13.1 (2022), pp. 1–11. DOI: 10.1186/s13326-022-00266-4. URL: https://hal.inria.fr/hal-03602177.

[17] ***Best Paper***
M. Melkonian, C. Juigné, O. Dameron, G. Rabut and E. Becker. 'Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases'. In: *Bioinformatics* (7th Jan. 2022), pp. 1–7. DOI: 10.1093/bioinformatics/btac013. URL: https://hal.archives-ouvertes.fr/hal-03522989.

[18] ***Best Paper***
A. Niarakis, D. Waltemath, J. Glazier, F. Schreiber, S. Keating, D. Nickerson, C. Chaouiya, A. Siegel, V. Noël, H. Hermjakob, T. Helikar, S. Soliman and L. Calzone. 'Addressing barriers in comprehensiveness, accessibility, reusability, interoperability and reproducibility of computational models in systems biology'. In: *Briefings in Bioinformatics* 23.4 (8th June 2022), pp. 1–11. DOI: 10.1093/bib/bbac212. URL: https://hal.inria.fr/hal-03690604.

[19] F. G. Petit, S. P. Jamin, P.-Y. Kernanec, E. Becker, G. Halet and M. Primig. 'EXOSC10/Rrp6 is essential for the eight-cell embryo/morula transition'. In: *Developmental Biology* 483 (2022), pp. 58–65. DOI: 10.1016/j.ydbio.2021.12.010. URL: https://hal.science/hal-03713237.

[20] M. Popova, I. Fakih, E. Forano, A. Siegel, R. Muñoz-Tamayo and D. Morgavi. 'Rumen microbial genomics: from cells to genes (and back to cells)'. In: *CAB Reviews Perspectives in Agriculture Veterinary Science Nutrition and Natural Resources* 2022 (19th Aug. 2022). DOI: 10.1079/cabireviews202217025. URL: https://hal.inrae.fr/hal-03929845.

[21] K. Thuillier, C. Baroukh, A. Bockmayr, L. Cottret, L. Paulevé and A. Siegel. 'MERRIN: MEtabolic Regulation Rule INference from time series data'. In: *Bioinformatics* 38.Supplement_2 (2022), pp. ii127–ii133. DOI: 10.1093/bioinformatics/btac479. URL: https://hal.archives-ouvertes.fr/hal-03701755.

[22] P. Vignet, J. Coquet, S. Aubert, M. Boudet, A. Siegel and N. Théret. 'Discrete modeling for integration and analysis of large-scale signaling networks'. In: *PLoS Computational Biology* (2022). DOI: 10.1371/journal.pcbi.1010175. URL: https://hal.inria.fr/hal-03693653.

International peer-reviewed conferences

[23] O. Abdou Arbi, A. Siegel and J. Bourdon. 'Contributions des entrées sur les sorties pour les réseaux métaboliques sur génomes entiers: performances et utilisation pour des études en nutrition humaine'. In: CARI 2022 - Colloque Africain sur la Recherche en Informatique et en Mathématiques Appliquées. Yaoundé, Cameroon, 4th Oct. 2022, pp. 1–12. URL: https://hal.inria.fr/hal-03689107.

[24] V. Kmetzsch, E. Becker, D. Saracino, V. Anquetil, D. Rinaldi, A. Camuzat, T. Gareau, I. Le Ber and O. Colliot. 'A multimodal variational autoencoder for estimating progression scores from imaging and microRNA data in rare neurodegenerative diseases'. In: SPIE Medical Imaging 2022: Image Processing. Vol. 12032. Medical Imaging 2022: Image Processing. San Diego, California, United States, 20th Feb. 2022, pp. 376–382. DOI: 10.1117/12.2607250. URL: https://hal.archives-ouvertes.fr/hal-03576117.

Conferences without proceedings

[25] O. Dameron, Y. Tirlet, M. Boudet and F. Legeai. 'Intégration de données de phénotypiques, environnementales et de biodiversité à l'aide des technologies du Web Sémantique'. In: SYSINFO-INRAe 2022 - Journées INRAE systèmes d'information pour les données agro-environnementales. Clermont-Ferrand, France, 21st Nov. 2022. URL: https://hal.inria.fr/hal-03893738.

[26] C. Juigné, O. Dameron, F. Gondret and E. Becker. 'A method to identify target molecules and extract the corresponding graph of interactions in BioPAX'. In: BBCC2022 - Bioinformatics and Computational Biology Conference. Virtual, Italy, 13th Dec. 2022. URL: https://hal.inria.fr/hal-03876091.

[27] ***Best Paper***
C. Juigné, O. Dameron, F. Moreews, F. Gondret and E. Becker. 'Detection and correction of non-conformities and redundancies in complexes of molecules in BioPAX'. In: Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM). Rennes, France, 5th July 2022, pp. 1–25. URL: https://hal.inrae.fr/hal-03752473.

[28] V. Kmetzsch, E. Becker, D. Saracino, V. Anquetil, D. Rinaldi, A. Camuzat, T. Gareau, I. Le Ber and O. Colliot. 'Highlight on Computing disease progression scores using multimodal variational autoencoders trained with neuroimaging and microRNA data'. In: Jobim 2022 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Rennes, France, 5th July 2022. URL: https://hal.science/hal-03877191.

[29] ***Best Paper***
M. Melkonian, C. Juigné, O. Dameron, G. Rabut and E. Becker. 'Highlight on semantic web technologies are effective to remove redundancies from protein-protein interaction databases and define reproducible interactomes'. In: Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM). Vol. 38. Proceedings. Rennes, France, 7th Jan. 2022, p. 146. DOI: 10.1093/bioinformatics/btac013. URL: https://hal.science/hal-03877219.

Scientific book chapters

[30] M. Bouguéon, P. Boutillier, J. Feret, O. Hazard and N. Théret. 'The rule-based model approach. A Kappa model for hepatic stellate cells activation by TGFB1'. In: *Systems Biology Modelling and Analysis: Formal Bioinformatics Methods and Tools*. Wiley, Nov. 2022, pp. 1–76. URL: https://hal.inria.fr/hal-03388100.

[31] O. Dameron. 'Méthodes du Web sémantique pour l'intégration de données en sciences de la vie'. In: *Intégration de données biologiques*. ISTE Group, July 2022, pp. 1–30. URL: https://hal.inria.fr/hal-03720874.

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

Doctoral dissertations and habilitation theses

[33] E. Becker. 'From homogeneous data to heterogeneous data in systems biology'. Université de Rennes 1, 14th Dec. 2022. URL: https://hal.archives-ouvertes.fr/tel-03906598.

[34] A. Belcour. 'Combining knowledge-based and sequence comparison approaches to elucidate metabolic functions, from pathways to communities'. Université de Rennes 1 (UR1), Rennes, FRA., 21st Oct. 2022. URL: https://theses.hal.science/tel-03924107.

[35] O. Dennler. 'Characterization in functional modules of ADAMTS-TSL proteins, by phylogeny approaches'. Université Rennes 1, 19th Dec. 2022. URL: https://hal.science/tel-03927428.

[36] V. Kmetzsch. 'Multimodal analysis of neuroimaging and transcriptomic data in genetic frontotemporal dementia'. Sorbonne Universite, 26th Sept. 2022. URL: https://theses.hal.science/tel-03892615.

Reports & preprints

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

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

[39] I. Fakih, J. Got, C. E. Robles-Rodriguez, A. Siegel, E. Forano and R. Muñoz-Tamayo. *Dynamic genome-based metabolic modeling of the predominant cellulolytic rumen bacteriumFibrobacter succinogenesS85*. 25th Nov. 2022. DOI: 10.1101/2022.10.18.512662. URL: https://hal.inrae.fr/hal-03871397.

Other scientific publications

[40] T. Antoine and F. Coste. 'Pondération d'automates obtenus par alignements partiels de séquences protéiques'. Ecole normale supérieure de Rennes - ENS Rennes, 30th Aug. 2022. URL: https://hal.inria.fr/hal-03789182.

[41] M. Aulagner. 'Draft reconstructions of metabolic networks from Nanopore and shotgun metagenomic data.' Université de Rennes 1, 1st July 2022. URL: https://hal.inria.fr/hal-03879996.

[42] C. Beust. 'Search for exposomic causality of liver fibrosis using network analysis'. Université de Rennes 1, France, 4th July 2022. URL: https://hal.inria.fr/hal-03936209.

[43] N. Buton, Y. Le Cunff and F. Coste. 'EnzBert: Deep attention network for enzyme class predictions'. In: JOBIM 2022 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Rennes, France, 5th July 2022, pp. 1–1. URL: https://hal.archives-ouvertes.fr/hal-03780557.

[44] O. Dennler, S. Blanquart, F. Coste, C. Belleannée and N. Théret. 'Functional Motif Prediction in ADAMTS-TSL proteins Based on Module(s) and Phenotype(s) Co-appearance'. In: ALPHY/AIEM 2022 - Rencontres Génomique évolutive, Bioinformatique, Alignement et Phylogénie. Rennes, France, 14th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-03870590.

[45] P. Hamon-Giraud. 'Comparative analysis of genome scale metabolic networks in brown algae'. Université Rennes 1, 15th June 2022. URL: https://hal.inria.fr/hal-03870140.

[46] C. Juigné. 'Integration and analysis of heterogeneous biological data modelled with multilayer graphs for a better understanding of feed efficiency'. In: Séminaire DIGIT-BIO INRAE. Ecully, France, 8th Dec. 2022. URL: https://hal.inria.fr/hal-03880428.

[47] H. Talibart, F. Coste and M. Carpentier. 'PPalign: optimal alignment of Potts models representing proteins with direct coupling information'. In: ISMB 2022 - 30th Conference on Intelligent Systems for Molecular Biology. Madison, United States, 10th July 2022, pp. 1–1. URL: https://hal.inria.fr/hal-03926272.

[48] Y. Tirlet. 'Integration of farming and environmental data thanks to Web semantic technologies'. Université de Rennes 1, 15th June 2022. URL: https://hal.inria.fr/hal-03870109.

11.3 Cited publications

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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