IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Activity Report 2016

Project-Team DYLISS

Dynamics, Logics and Inference for biological Systems and Sequences

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

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Computational Biology

Table of contents

Project-Team DYLISS

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

Keywords:

Computer Science and Digital Science:

3.1.1. - Modeling, representation
3.2.3. - Inference
3.2.4. - Semantic Web
3.2.5. - Ontologies
3.3. - Data and knowledge analysis
7.2. - Discrete mathematics, combinatorics
7.3. - Optimization
7.4. - Logic in Computer Science
8.1. - Knowledge
8.2. - Machine learning
8.7. - AI algorithmics

Other Research Topics and Application Domains:

1.1.2. - Molecular biology
1.1.3. - Cellular biology
1.1.9. - Bioinformatics
1.1.11. - Systems biology
1.1.14. - Microbiology
2.2.3. - Cancer
2.2.5. - Immune system diseases

1. Members

Research Scientists

François Coste [Inria, Researcher]
Jacques Nicolas [Inria, Senior Researcher, HDR]
Anne Siegel [Team leader, CNRS, Senior Researcher, HDR]

Faculty Members

Catherine Belleannée [Univ. Rennes I, Associate Professor, 90%-time presence in the lab]
Olivier Dameron [Univ. Rennes I, Associate Professor, HDR]
Laurent Miclet [Univ. Rennes I, Professor, until Sep 2016, HDR]

Engineers

Jeanne Got [CNRS]
Meziane Aite [Inria, ADT]
Marie Chevallier [CNRS, ADT until Sept 2016, Regional contract since Oct 2016]
Xavier Garnier [INRA]
Camille Trottier [CNRS, ANR IDEALG]
Pierre Vignet [CNRS, since Nov 2016, ITMO Cancer TGFSysBio]

PhD Students

Aymeric Antoine-Lorquin [Inria]

Lucas Bourneuf [Univ. Rennes I, since Oct 2016]
Jean Coquet [Univ. Rennes I]
Victorien Delannée [Univ. Rennes I]
Clémence Frioux [Inria]
Julie Laniau [Inria]
Yann Rivault [Univ. Rennes I]

Post-Doctoral Fellow
Aurélie Evrard [INRA, until Aug 2016]

Visiting Scientists
Oumarou Abdou Arbi [Univ. of Maradi - Niger, from Jan 2016 until Feb 2016]
Maria Paz Cortes Burgos [Univ. of Chile, from may 2016 until Nov 2016]
Mauricio Alejandro Latorre [Univ. of Chile, Sep 2016]
Alejandro Maass [Univ. of Chile, Sep 2016]
Santiago Videla [Institut Leloir, Argentina, Jul 2016]

Administrative Assistant
Marie Le Roic [Univ. Rennes I]

Others
Jérémie Bourdon [Univ. Nantes, Dyliss associate member, 20%-time presence in the lab, HDR]
Damien Eveillard [Univ. Nantes, Dyliss associate member, 20%-time presence in the lab]
François Moreews [INRA, Dyliss partner, 80%-time presence in the lab]
Nathalie Théret [INSERM, Senior Researcher, Dyliss partner, 80%-time presence in the lab, HDR]
Denis Tagu [INRA, Senior Researcher, Dyliss associate member, 20%-time presence in the lab, HDR]

2. Overall Objectives

2.1. Overall objectives

The research domain of the bioinformatics Dyliss team is sequence analysis and systems biology. Our main goal in biology is to characterize groups of genetic actors that control the phenotypic answer of species when challenged by their environment. The team explores methods in the field of formal systems, more precisely in knowledge representation, constraints programming, multi-scale analysis of dynamical systems, and machine learning. Our goal is to identify key regulators of the environmental response by structuring and reasoning on information which combines physiological responses measured with omics technologies (RNA-seq, metabolomics, proteomics), genetic information from their long-distant cousins and knowledge about regulation and metabolic pathways stored in public repositories.

The main challenges we face are data incompleteness and heterogeneity. We favor the construction and study of a "space of feasible models or hypotheses" including known constraints and facts on a living system *rather than searching for a single optimized model*. We develop methods allowing a precise investigation of this space of hypotheses. Therefore, we are in position of developing experimental strategies to progressively shrink the space of hypotheses and gain in the understanding of the system. Importantly, one should notice that our models spans a quite large spectrum of discrete structures: oriented graphs, boolean networks, automata, or expressive grammars.

More concretely, the steps of the analysis are to (*i*) formalize and integrate in a set of logical or grammatical constraints both generic knowledge information (litterature-based regulatory pathways, diversity of molecular functions, DNA patterns associated with molecular mechanisms) and species-specific information (physiological response to perturbation, sequencing...); (*ii*) investigate the space of admissible models and exhibit its main features by solving combinatorial optimization problems; (*iii*) identify corresponding genomic products within sequences. At each of these steps, we rely on symbolic methods for model space exploration: ontologies and formal concepts analysis.

We target applications for which large-scale heterogeneous data about a specific but complex physiological phenotype are available. Existing long-term partnerships with biological labs give strong support to this choice. In marine biology, we collaborate closely with the Station biologique de Roscoff (*Idealg*, Investissement avenir "Bioressources et Biotechnologies"). In environmental microbiology we collaborate both with the CRG in Chile in the framework of the Ciric Chilean Inria center (*Ciric-Omics*). In agriculture, our main partners are within the INRA institute in Rennes, with a focus on the understanding of pea-aphids microbiology and of breeding animals metabolism (porc, chicken, cow). More recently, we have introduced health as a new application field of the team, especially through the study of large-scale boolean networks and their confrontation with knowledge repositories.

3. Research Program

3.1. Modeling knowledge integration with combinatorial constraints

Biological networks are built with data-driven approaches aiming at translating genomic information into a functional map. Most methods are based on a probabilistic framework which defines a probability distribution over the set of models. The reconstructed network is then defined as the most likely model given the data.

Our team has investigated an alternative perspective where each data induces a set of constraints - related to the steady state response of the system dynamics - on the set of possible values in a network of fixed topology. The methods that we have developed complete the network with product states at the level of nodes and influence types at the level of edges, able to globally explain experimental data. In other words, the selection of relevant information in the model is no more performed by selecting *the* network with the highest score, but rather by exploring the complete space of models satisfying constraints on the possible dynamics supported by prior knowledge and observations. In the (common) case when there is no model satisfying all the constraints, we relax the problem by introducing new combinatorial optimization problems that introduce the possibility of correcting the data or the knowledge. Common properties to all solutions are considered as a robust information about the system, as they are independent from the choice of a single solution to the optimization problem [6].

Solving these computational issues requires addressing NP-hard qualitative (non-temporal) issues. We have developed a long-term collaboration with Potsdam University in order to use a logical paradigm named **Answer Set Programming** (ASP) [50], [69] to solve these constraint satisfiability and combinatorial optimization issues. Applied on transcriptomic or cancer networks, our methods identified which regions of a large-scale network shall be corrected [51], and proposed robust corrections [5]. This result suggested that this approach was compatible with efficiency, scale and expressivity needed by biological systems.

During the last years, our goal was to provide **formal models of queries on biological networks** with the focus of integrating dynamical information as explicit logical constraints in the modeling process. Using these technologies requires to revisit and reformulate constraint-satisfiability problems at hand in order both to decrease the search space size in the grounding part of the process and to improve the exploration of this search space in the solving part of the process. Concretely, getting logical encoding for the optimization problems forces to clarify the roles and dependencies between parameters involved in the problem. This paves the way to a refinement approach based on a fine investigation of the space of hypotheses in order to make it smaller and gain in the understanding of the system. Our studies confirmed that logical paradigms are a powerful approach to build and query reconstructed biological systems, in complement to discriminative ("black-box") approaches based on statistical machine-learning. Based on these technologies, we have developed a panel of methods allowing the integration of muli-scale data knowledge, linking genomics, metabolomics, expression data and protein measurement of several phenotypes (see Fig. 1).

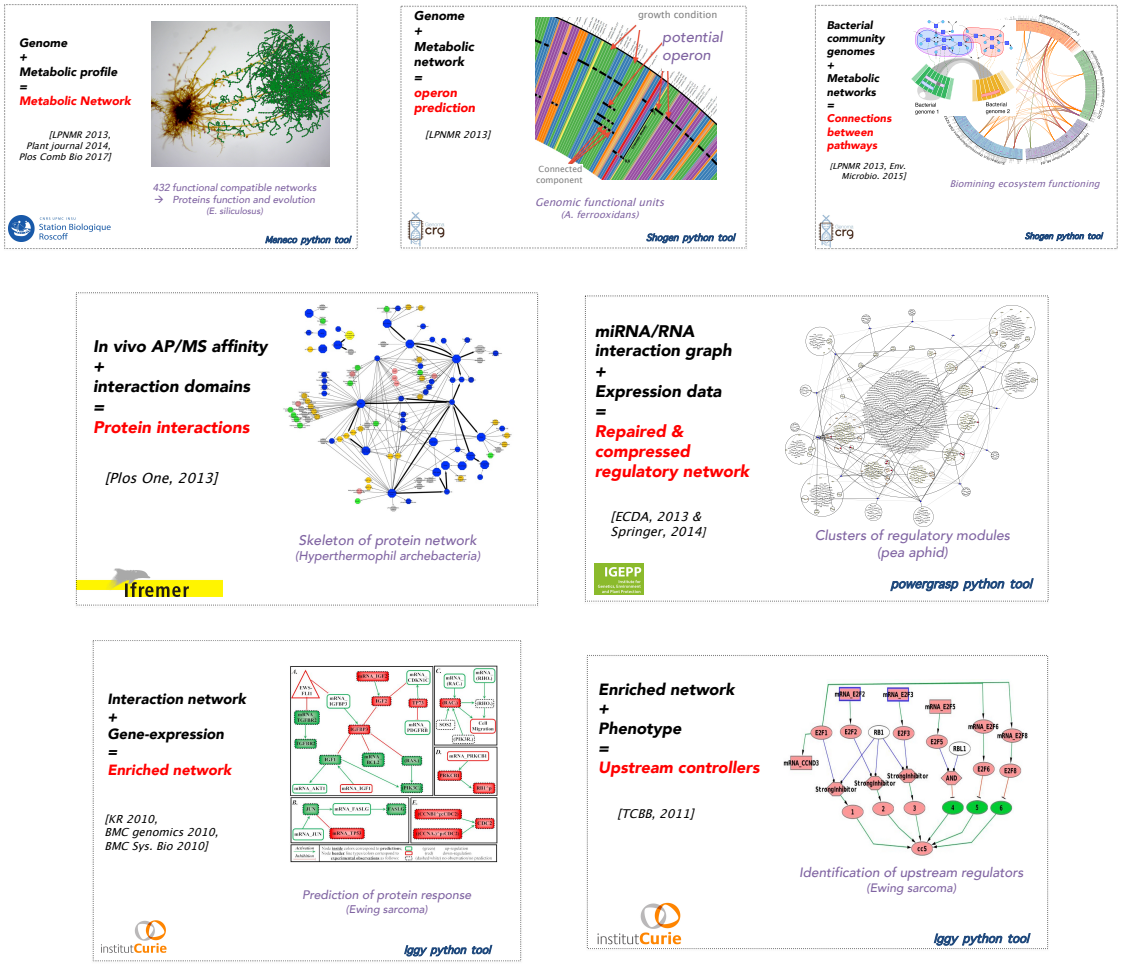

Figure 1. ***Multi-scale data and knowledge integration procedures*** *Dynamical analyses are undergone to elucidate relationships between biological scales. Such dependencies are combined with data and turn into a first order logic paradigm (Answer Set Programming). Key interactions or genes of interest are then identified to be the solution of a combinatorial optimization problem. Methods are encapsulated in python packages and provided in the meta-package bioasp.*

Notice that our main issue is in the field of knowledge representation. More precisely, we do not wish to develop new solvers or grounders, a self-contained computational issue which is addressed by specialized teams such as our collaborator team in Potsdam. Our goal is rather to investigate how the constant progresses in the field of constraint logical programming, shown by the performance of ASP-solvers, are sufficient to address the complexity of constraint-satisfiability and combinatorial optimization issues explored in systems biology. In this direction, we work in close interaction with Potsdam university to feed their research activities which challenging issues from bioinformatics and, as a feed-back, take benefit of the prototypes they develop.

By exploring the complete space of models, our approach typically produces numerous candidate models compatible with the observations. We began investigating to what extent domain knowledge can further refine the analysis of the set of models by identifying classes of similar models, or by selecting a subset of models that satify an additional constraint (for instance, best fit with a set of experiments, or with a minimal size). We anticipate that this will be particularly relevant when studying non-model species for which little is known but valuable information from other species can be transposed or adapted. These efforts consist in developing reasoning methods based on ontologies as formal representation of symbolic knowledge. We use Semantic Web tools such as SPARQL for querying and integrating large sources of external knowledge, and measures of semantic similarity and particularity for analyzing data.

3.2. Modeling the dynamical response of biological systems with logical and (non)-linear constraints

As explained below, Answer Set programming technologies enable the identification of key controllers based on the integration of static data. As a natural follow-up, we also develop optimization techniques to learn models of the dynamics of a biological system. As before, our strategy is not to select a single model fitting with experimental data but rather to decipher the complete set of families of models which a compatible with the observed response. Our main research line in this field is to decipher the appropriate level of expressivity (in terms of constraints) allowing both to properly report the nature of data and knowledge and to allow for an exhaustive study of the space of feasible models. To implement this strategy, we rely on several constraint programming frameworks, which depend on the model scale and the nature of time-points kinetic measurements. The three following examples are shown in Fig. 2.

- In [7], logical programming (Answer Set programming) is used to decipher the combinatorics ot synchrone boolean networks explaining static or dynamics response of signaling networks to perturbations (such as measured by phosphoproteomics technologies).
- In [49], SAT-based approaches are used to decipher the combinatorics of large-scale asynchronous boolean networks. In order to gain in expressivity, we model these networks as guarded-transition network, an extension of Petri nets.
- In [2] and [47], linear Programming frameworks are used to decipher the variability of the response of reaction-based networks. Still to gain in expressivity, we model systems with Markovian qualitative description of its dynamics together with quantitative laws which describe the effect of the dynamic transitions over higher scale quantitative measurements. Families of models are investigated with ad-hoc local search algorithms.
- Finally, classical learning methods are used to build ad-hoc parameterized numerical models that provide the most parsimonious explanations to experimental measurements.

3.3. Modeling sequences with formal grammars

Once groups of genome products implied in the answer of the species have been identified with integrative or dynamics methods, it remains to characterize the biological actors within genomes. To that goal, we both learn, model and parse formal patterns within DNA, RNA or protein sequences. More precisely, our research on modeling biomolecular sequences with expressive formal grammars focuses on learning such grammars from examples, helping biologists to design their own grammar and providing practical parsing tools.

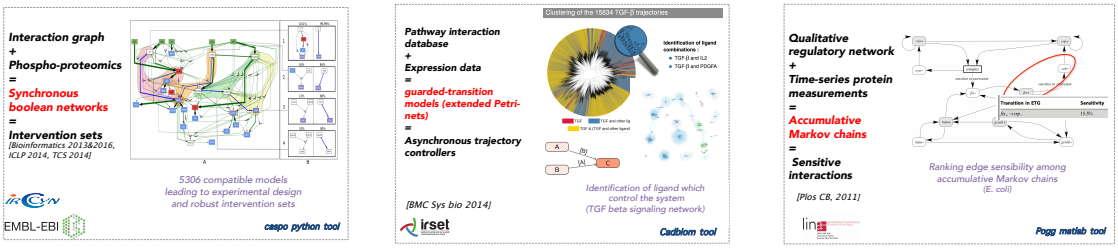

Figure 2. ***Learning and investigating complete families of dynamical models compatible with available data.*** *Depending on the scale of the system and the nature of data, we use synchronous boolean networks, enriched Petri Nets or accumulative Markov chains to report and explain the measured response of a biological systems.*

On the development of **machine learning** algorithms for the induction of grammatical models [40], we have a strong expertise on learning finite state automata. We have proposed an algorithm that learns successfully automata modeling families of (non homologous) functional families of proteins [4], leading to a tool named Protomata-learner (see Fig. 3). The algorthim is based on a similar fragment merging heuristic approach which reports partial and local alignments contained in a family of sequences.. As an example, this tool allowed us to properly model the TNF protein family, a difficult task for classical probabilistic-based approaches. It was also applied successfully to model important enzymatic families of proteins in cyanobacteria [3]. Our future goal is to further demonstrate the relevance of formal language modeling by addressing the question of a fully automatic prediction from the sequence of all the enzymatic families, aiming at improving even more the sensitivity and specificity of the models. As enzyme-substrate interactions are very specific central relations for integrated genome/metabolome studies and are characterized by faint signatures, we shall rely on models for active sites involved in cellular regulation or catalysis mechanisms. This requires to build models gathering both structural and sequence information in order to describe (potentially nested or crossing) long-term dependencies such as contacts of amino-acids that are far in the sequence but close in the 3D protein folding. Our current researches is focused on the inference of Context-Free Grammars including the topological information coming from the structural characterization of active sites.

Using context-free grammars instead of regular patterns increases the complexity of **parsing** issues. Indeed, efficient parsing tools have been developed to identify patterns within genomes but most of them are restricted to simple regular patterns. Definite Clause Grammars (DCG), a particular form of logical context-free grammars have been used in various works to model DNA sequence features [76]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 3) [90], [89]. This increases the expressivity of the formalism towards mildly context sensitive grammars. Thus, those grammars model not only DNA/RNA sequence features but also structural features such as repeats, palindromes, stem/loop or pseudo-knots. Few years ago, we have designed a first tool, STAN (suffix-tree analyser), in order to make it possible to search for a subset of SVG patterns in full chromosome sequences [8]. This tool was used for the recognition of transposable elements in *Arabidopsis thaliana* [92]. We have enlarged this experience through a new modeling language, called Logol [1]. Generally, a suitable language for the search of particular components in languages has to meet several needs : expressing existing structures in a compact way, using existing databases of motifs, helping the description of interacting components. In other words, the difficulty is to find a good tradeoff between expressivity and complexity to allow the specification of realistic models at genome scale. The Logol language and associated framework have been built in this direction. See Figure 3 for illustration. The Logol specificity beside other SVG-like languages mainly lies in a systematic introduction of constraints on string variables.

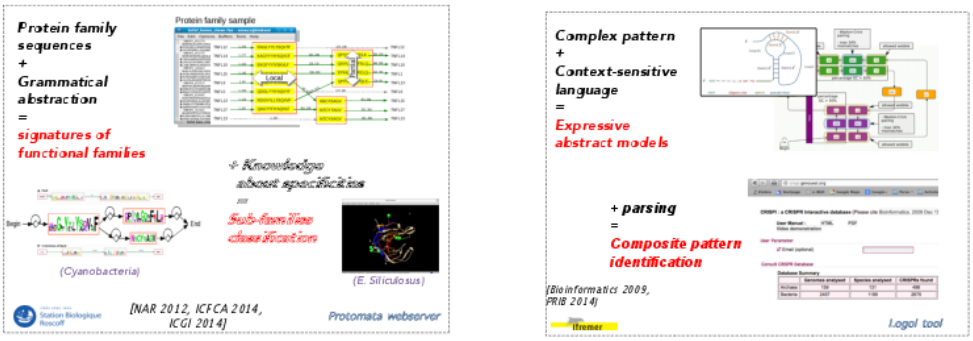

Figure 3. ***Learning and parsing sequences of genome or protein families with expressive grammars****. (a) The protomata workflow starts from a set of protein sequences. A partial local alignment is computed and an automaton is inferred, which can be considered as a signature of the family of proteins. This allows searching for new members of the family [3]. Adding further information about the specific properties of proteins within the family allows to exhibit a refined classification. (b) The Logol framework allows modeling complex structure in sequences, such as a pseudo-knot (RNA structure). This is based on the expressivity of String Variable Grammars. Combined with parsers, this leads to composite pattern identification such as CRISPR. [84].*

3.4. Symbolic methods for model space exploration: Semantic web for life sciences and Formal Concepts Analysis

All the methods presented in the previous sections usually result in pools of candidates which equivalently explain the data and knowledge. These candidates can be dynamical systems, compounds, biological sequences, proteins... In any case, the output of our formal methods generally requires *a posteriori* investigation and filtering by domain experts. In order to assist them, we rely on two classes of symbolic technics: Semantic Web technologies and Formal Concept Analysis (FCA). They both aim at the formalization and management of knowledge, that is, the explicitation of relations occuring in structured data. These technics complement each other: the production of relevant concepts in FCA highly depends on the availability of semantic annotations using a controlled set of terms and conversely, building and exploiting ontologies is a complex process that can be made much easier with FCA.

Integrating heterogenous data with semantic web technologies The emergence of ontologies in biomedical informatics and bioinformatics happened in parallel with the development of the **Semantic Web** in the computer science community [88]. Let us recall that the Semantic Web is an extension of the current Web that provides an infrastructure integrating data and ontologies in order to support unified reasoning. Since the beginning, life sciences have been a major application domain for the Semantic Web [52]. This was motivated by the joint evolution of data acquisition capabilities in the biomedical field, and of the methods and infrastructures supporting data analysis (grids, the Internet...), resulting in an explosion of data production in complementary domains [60], [53]. Consequently, Semantic Web technologies have become an integral part of translational medicine and translational bioinformatics [63]. The Linked Open Data project promotes the integration of data sources in machine-processable formats compatible with the Semantic Web [59], with a strong involvement of life sciences in this initiative.

However, a specificity of life sciences "data deluge" is that the proportion of generated data is much higher than in the more general "big data phenomenon", and that these data are highly connected [91]. **The bottleneck that once was data scarcity now lies in the lack of adequate methods supporting data integration, processing and analysis**. [78]. Each of these steps typically hinges on domain knowledge, which is why they resist automation. This knowledge can be seen as the set of rules representing in what conditions data can be used or can be combined for inferring new data or new links between data.

In this setting, we are working on the integration of Semantic Web resources with our data analysis methods in order to take existing biological knowledge into account. We have introduced several methods to interpret semantic similarities and particularities [58], [57]. We now focus our attention on the semi-automated construction of RDF abstractions of heterogeneous datasets which can be handled by non-expert users. This allows both to automatically prepare input datasets for the other methods developed in the team and to analyse the output of the methods in a wide knowledge context.

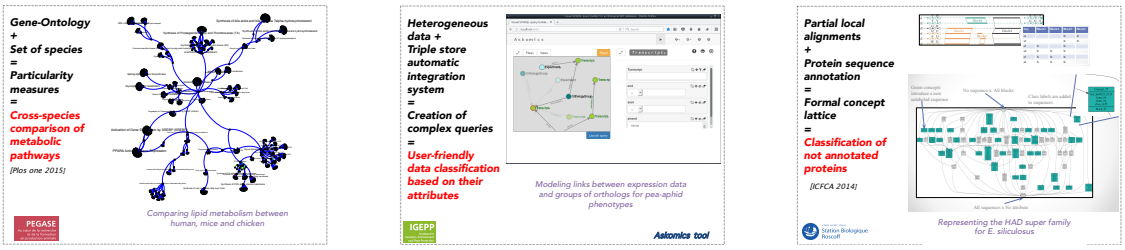

Figure 4. ***Data-sciences methods based-on semantic-web technologies and formal concept analysis*** *allows for the knowledge-based post-processing of the results of bioinformatics methods.*

Using Formal concept analysis to explore the results of bioinformatics analyses Formal concept analysis aims at the development of conceptual structures which can be logically activated for the formation of judgments and conclusions [96]. It is used in various domains managing structured data such as knowledge processing, information retrieval or classification [79]. In its most simple form, one considers a binary relation between a set of objects and a set of attributes. In this setting, formal concept analysis formalizes the semantic notions of extension and intension. Concepts are related within a lattice structure (Galois connection) by subconcept-superconcept relations, and this allows drawing causality relations between attribute subsets. In bioinformatics, it has been used to derive phylogenetic relations among groups of organisms [77], a classification task that requires to take into account many-valued Galois connections. We have proposed in a similar way a classification scheme for the problem of protein assignment in a set of protein families [65].

One of the most important issue with concept analysis is due to the fact that current methods remain very sensitive to the presence of uncertainty or incompleteness in data. On the other hand, this apparent defect can be reversed to serve as a marker of incompleteness or inconsistency [66]. Following this inspiration, we have proposed a methodology to tackle the problem of uncertainty on biological networks where edges are mostly predicted links with a high level of false positives [97]. The general idea consists to look for a tradeoff between the simplicity of the conceptual representation and the need to manage exceptions. As a very prospective challenge, we are exploring the idea of using ontologies to help this or to help ontology refinement using concept analysis [80], [56], [83].

More generally, common difficult tasks in this context are visualization, search for local structures (graph mining) and network comparison. Network compression is a good solution for an efficient treatment of all these tasks. This has been used with success in power graphs, which are abstract graphs where nodes are clusters of nodes in the initial graph and edges represent bicliques between two sets of nodes [85]. In fact, concepts are maximal bicliques and we are currently developing the power graph idea in the framework of concept analysis.

4. Application Domains

4.1. Application domain in bioinformatics

As mentioned before, our main goal in biology is to characterize groups of genetic actors that control the response of living species capable of facing extreme environments. To focus our developments, applications and collaborations, we have identified three biological questions which deserve integrative studies. Each axis may be considered independently from the others although their combination, a mid-term challenge, will have the best impact in practice towards the long-term perspective of identifying proteins controlling the production of a metabolite of industrial interest. It is illustrated in our presentation for a major algae product: polyunsaturated fatty acids (PUFAs) and their derivatives.

Integrative biology with combinatorial optimization. The first axis of the project (data integration) aims at identifying *who* is involved in the specific response of a biological system to an environmental stress. Targeted actors will mainly consist in groups of genetic products or biological pathways. For instance, which pathways are involved in the specific production of PUFAs in brown algae? The main work is to represent in a system of logical constraints the full knowledge at hand concerning the genetic or metabolic actors, the available observations and the effects of the system dynamics. To this aim, we focus on the use of Answer Set Programming as we are experienced in modeling with this paradigm and we have a strong partnership with a computer science team leader in the development of dedicated grounders and solvers (Potsdam university). See Sec. 3.1.

Systems biology with discrete dynamical modeling. Once a model is built and its main actors are identified, the next step is to clarify *how* they combine to control the system. This is the second axis of the project. Roughly, the fine tuning of the system response may be of two types. Either it results from the discrete combinatorics of the actors, as the result of a genetic adaptation to extreme environmental conditions or the difference between species is rather at the enzyme-efficiency level. For instance, if PUFAs are found to be

produced using a set of pathways specific to brown algae, our work on dynamical modeling will consist to apply constraint-based combinatorial approaches to select consistent combinations of pathways controlling the metabolite production. Otherwise, if enzymes controlling the production of PUFAs are found to be expressed in other algaes, it suggests that the response of the system is rather governed by a fine quantitative tuning of pathways. In this case, we use symbolic dynamics and average-case analysis of algorithms to weight the respective importance of interactions in observed phenotypes (see Sec. 3.2 and Fig. 2). This specific approach is motivated by the quite restricted spectrum of available physiological observations over the asymptotic dynamics of the biological system.

Biological sequence annotation with grammatical inference and modelling In order to check the accuracy of in-silico predictions, a third research axis of the team is to extract genetic actors responsible of biological pathways of interest in the targeted organism and locate them in the genome. In our guiding example, active proteins implied in PUFAs controlling pathways have to be precisely identified. Actors structures are represented by syntactic models (see Fig. 3). We use knowledge-based induction on far instances for the recognition of new members of a given sequence family within non-model genomes (see Fig. 3). A main objective is to model enzyme specificity with highly expressive syntactic structures - context-free model - in order to take into account constraints imposed by local domains or long-distance interactions within a protein sequence. See Sec. 3.3 for details.

Data classification with data sciences All the methods presented in the previous section usually result in pools of candidates which equivalently explain the data and knowlegde. These candidates can by dynamical systems, compounds, biological sequences, proteins... In any case, the output of our formal methods generally deserves a a-posteriori investigation and filtering. To that goal, we rely on two classes of symbolic technics: semantic web technologies and Formal Concept Analysis See Sec. 3.4 for details.

4.2. Application fields in biology

Our methods are applied in several fields of molecular biology.

Our main application field is **marine biology**, as it is a transversal field with respect to issues in integrative biology, dynamical systems and sequence analysis. Our main collaborators work at the Station Biologique de Roscoff. We are strongly involved in the study of brown algae: the *meneco, memap and memerge* tools were designed to realize a complete reconstruction of metabolic networks for non-benchmark species [82], [68]. On the same application model, the pattern discovery tool *protomata learner* combined with supervised bi-clustering based on formal concept analysis allows for the classification of sub-families of specific proteins [65]. The same tool also allowed us to gain a better understanding of cyanobacteria proteins [3]. At the larger level of 4D structures, classification technics have also allowed us to introduce new methods for the characterization of viruses in marine metagenomic sample [19]. Finally, in dynamical systems, we use asymptotic analysis (tool *pogg*) to decipher the initiation of sea urchin translation [55] [24]. We are currently involved in two new applications in this domain: the team participates to a Inria Project Lab program with the Biocore and Ange Inria teams, focused on the understanding on green micro-algae; and we are involved in the deciphering of phytoplancton variability at the system biology level in collaboration with the Station Biologique de Roscoff (ANR Samosa).

In **micro-biology**, our main issue is the understanding of bacteria living in extreme environments, mainly in collaboration with the group of bioinformatics at Universidad de Chile (funded by CMM, CRG and Inria-Chile). In order to elucidate the main characteristics of these bacteria, we develop efficient methods to identify the main groups of regulators for their specific response in their living environment. To that purpose, we use constraints-based modeling and combinatorial optimization. The integrative biology tools *meneco bioquali, ingranalysis, shogen, lombarde* were designed in this context [6]. in 2016, two applications focused on the study of extremophile consortium of bacteria have been performed with these tools [15], [13]. In parallel, in collaboration with Ifremer (Brest), we have conducted similar work to decipher protein-protein interactions within archebacteria [81]. Our sequence analysis tool (*logol*) allowed us to build and maintain a very expressive CRISPR database [10] [54].

Similarly, in **agriculture**, our goal is to propose methods to identify regulators of very complex phenotypes related to environmental issues. In collaboration with researchers from Inra/Pegase and Inra/Igeep laboratories, we develop methods to distinguish the response of breeding animals to different diaries or treatments [47] and characterize upstream transcriptional regulators [61], with applications in porks [70], [71] [20]. The pattern matching tool *logol* also allows for a fine identification of transcription factor motifs applied to chicken [67] [54]. Semantic-based analysis was useful for interpreting differences of gene expression in pork meat [72]. Finally, Constraints-based programming also allows us to decipher regulators of reproduction for pea aphids [75], [98] and paved the way to the recent research track initiated in the team about integration of heterogeneous data with RDF-technologies (see askomics software) [37], [45].

Similarly, in **agriculture**, our goal is to propose methods to identify regulators of very complex phenotypes related to environmental issues. In collaboration with researchers from Inra/Pegase laboratory, we develop methods to distinguish the response of breeding animals to different diaries or treatments [47] and characterize upstream transcriptional regulators [61], applied to porks [70], [71] [20]. The pattern matching tool *logol* also allows for a fine identification of transcription factor motifs applied to chicken [67] [54]. Semantic-based analysis was useful for interpreting differences of gene expression in pork meat [72].

In addition, constraints-based programming also allows us to decipher regulators of reproduction for the pea aphid, an insect that is a pest on plants [75], [98]. This was performed in collaboration with Inra/Igepp . This paved the way to the recent research track initiated in the team about integration of heterogeneous data with RDF-technologies (see askomics software) [37], [45] and about graph-compression (see powergrasp software).

In **bio-medical applications**, we focus our attention on the confrontation of large-scale measurements with large-scale knowledge repositories about regulation pathways such as Transpath, PID or pathway commons. In collaboration with Institut Curie, we have studied the Ewing Sarcoma regulation network to test the capability of our tool *bioquali* to accurately correct and predict a large-scale network behavior [51]. Our ongoing studies in this field focus on the exhaustive learning of discrete dynamical networks matching with experimental data, as a case study for modeling experimental design with constraints-based approaches. To that purpose, we collaborate with J. Saez Rodriguez group at EBI [94] and N. Theret group at Inserm/Irset (Rennes) [49]. The dynamical system tools *caspo and cadbiom* were designed within these collaborations. Ongoing studies focus on the understanding of the metabolism of xenobiotics (mecagenotox program) and the filtering of sets of regulatory compounds within large-scale signaling network (TGFSysBio project).

5. Highlights of the Year

5.1. Highlights of the Year

The first main novelty in 2016 is the release of our first methods and tools based on semantic web technologies. These methods enable the pre-processing of heterogeneous data prior to their integration in the toolboxes developped by the team. Methods for the transparent integration and querying of heterogeneous data (AskOmics) as well as the user-friendly tracable reconstruction of metabolic networks (PADmet package) have been developped in collaboration with our main partners (INRA Rennes, University of Chile, Station biologique de Roscoff) to facilitate the comparison of phenotypes accross several species or several strains.

6. New Software and Platforms

6.1. AskOmics

KEYWORDS: RDF - SPARQL - Querying - Graph
FUNCTIONAL DESCRIPTION

AskOmics allows to load heterogeneous bioinformatics data (formatted as tabular files or directly in RDF) into a Triple Store system using a user-friendly web interface. AskOmics also provides an intuitive graph-based user interface supporting the creation of complex queries that currently require hours of manual searches across tens of spreadsheet files. The elements of interest selected in the graph are then automatically converted into a SPARQL query that is executed on the users' data.

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

6.2. PADMet

PortAble Database for Metabolism
KEYWORDS: Bioinformatics - Toolbox - Metabolic networks - Standardization
FUNCTIONAL DESCRIPTION

The PADMet package allows conciliating genomics and metabolic network information used to produce a genome-scale constraint-based metabolic model within a database that traces all the reconstruction process steps. It allows representing the metabolic model in the form of a Wiki containing all the used/traced information. Other standard outputs are made available with the package. The main concept underlying PADMet-Package is to provide solutions that ensure the consistency, the internal standardization and the reconciliation of the information used within any workflow that combines several tools involving metabolic networks reconstruction or analysis. The PADMet package is at the core of the AuReMe workflow, dedicated to the primary reconstruction of genome-scale metabolic networks from raw data. It allows the study of organisms for which few experimental data are available. Its main feature is to undergo the reconstruction of the metabolic network by combining several heterogeneous knowledge and data sources, including the information reported by several scaffold metabolic networks for cousin species.

- Partners: CNRS - Inria - Université de Rennes 1 - University of Chile.
- Contact: Meziane Aite
- https://gitlab.inria.fr/DYLISS/padmet-toolbox

6.3. PowerGrASP

Power Graph compression in ASP
KEYWORDS: Bioinformatics - Constraint-based Programming - Data visualization - Optimization - Decomposition - Graph - Graph visualization - Pattern extraction - Answer Set Programming - Formal concept analysis
FUNCTIONAL DESCRIPTION

Implementation of graph compression methods oriented toward visualization, and based on power graph analysis. The method relies of formal concept analysis and is implemented in the declarative langage Answer Set Programming. It is applied to regulatory networks currently produced in the domain of bioinformatics.

- Participants: Lucas Bourneuf, Jacques Nicolas
- Partners: Inria - Université de Rennes 1 - INRA.
- Contact: Lucas Bourneuf
- URL: http://github.com/aluriak/powergrasp

6.4. Platforms and toolboxes

A goal of the team is to facilitate interplays between tools for biological data analysis and integration. Our tools aim at guiding the user to progressively reduce the space of models (families of sequences of genes or proteins, families of keys actors involved in a system response, dynamical models) which are compatible with both knowledge and experimental observations.

Most of our tools are developed in collaboration with the GenOuest resource and data center hosted in the IRISA laboratory, including their computer facilities [more info]. It worths considering them into larger dedicated environments to benefit from the expertise of other research groups.

- The **BioShadock** repository allows one to share the different docker containers that we are developing [website].
- The **Inria chile Mobyle portal** gathers some of the tools that were developed in collaboration with Dyliss, such as meneco, shogen and lombarde [website].
- The **bioASP portal** gather most of ASP-based python packages that we are developping in collaboration with Potsdam university [website]
- The **GenOuest galaxy portal** now provides access to most tools for integrative biology and sequence annotation (access on demand).

6.4.1. AuReMe - Tracable reconstruction of metabolic networks

The toolbox **AuReMe** allows for the **Au**tomatic **Re**construction of **Me**tabolic networks based on the combination of multiple heteregeneous data and knowledge sources. Since 2016, the workflow has been made available as a Docker image to facilitate its distribution among the scientific community [web page].

- The **Model-management PADmet module** allows conciliating genomics and metabolic network information used to produce the metabolic model within a local database that traces all the reconstruction process steps and to connect software in the pipeline. This toolbox was completly redesigned in 2016. [package]
- The **meneco python package** allows filling the gaps of a metabolic network by using a qualitative approach to elaborate the biosynthetic capacities; the problem is viewed as a combinatorial optimization problem encoded in a Answer Set Programming Problem [87] [64]. [python package].
- The **shogen python package** allows aligning genome and metabolic network to identify genome units which contain a large density of genes coding for enzymes that regulate successive reactions of metabolic pathways; the problem is also encoded with an ASP program. [62]. [python package].
- The **Manual curation assistance PADmet module** allows for curating the reported metabolic networks and modify metadata [package].
- The **Wiki-export PADmet module** enables the export of the metabolic network and its functional genomic unit as a local wiki platform allowing the user-friendly investigation of the network together with the main steps used to reconstruct it. It was developed in 2016. [package].

6.4.2. Filtering interaction networks with graph-based optimization criteria

The goal is to offer a toolbox for the reconstruction of networks from genome, literature and large-scale observation data (expression data, metabolomics...) in order to elucidate the main regulators of an observed phenotype. Most of the optimization issues are addressed with Answer Set Programming.

- The **lombarde package** enables the filtering of transcription-factor/binding-site regulatory networks with mutual information reported by the response to environmental perturbations. The high level of false-positive interactions is filters according to graph-based criteria. Knowledge about regulatory modules such as operons or the output of the shogen package can be taken into account [48][13] [web server].
- The **KeyRegulatorFinder package** allows searching key regulators of lists of molecules (like metabolites, enzymes or genes) by taking advantage of knowledge databases in cell metabolism and signaling. The complete information is transcribed into a large-scale interaction graph which is filtered to report the most significant upstream regulators of the considered list of molecules [61] [package].
- The **powerGrasp python package** provides an implementation of graph compression methods oriented toward visualization, and based on power graph analysis. [package].

- The **iggy package** enables the repairing of an interaction graph with respect to expression data. It proposes a range of different operations for altering experimental data and/or a biological network in order to re-establish their mutual consistency, an indispensable prerequisite for automated prediction. For accomplishing repair and prediction, we take advantage of the distinguished modeling and reasoning capacities of Answer Set Programming. [5] [93] [Python package] [web server].

6.4.3. Caspo - Studying synchronous boolean networks

The **caspo** pipeline is dedicated to automated reasoning on logical signaling networks. The main underlying issue is that inherent experimental noise is considered, so that many different logical networks can be compatible with a set of experimental observations.

Software provides an easy to use software for the study of synchronous logical (boolean) networks. In 2016, the tool was redesigned to enhance its functionalities and integrated in a docker container to facilitate its use on any platform [86] [28] [python package and docker container].

- The **caspo-learn module** performs an automated inference of logical networks from the observed response to different perturbations (phosphoproteomics datasets). It allows for identifying admissible large-scale logic models saving a lot of efforts and without any a priori bias. It is also included in the cellNopt package [1] [7] [94].
- The **caspo-classify, predict and visualize modules** allows for classifying a family of boolean networks with respect to their input–out- put predictions [7].
- The **caspo-design module** designs experimental perturbations which would allow for an optimal discrimination of rival models in a family of boolean networks [95].
- The **caspo-control module** identifies key-players of a family of networks: it computes robust intervention strategies (i.e. inclusion minimal sets of knock-ins and knock-outs) that force a set of target species or compounds into a desired steady state [73].
- **caspo-timeseries module** have been designed by our colleagues from LRI as an extension of the caspo pipeline to take into account time-series observation datasets in the learning procedure [23] [python package and docker container].

6.4.4. cadbiom - Building and analyzing the asynchronous dynamics of enriched logical networks

Based on Guarded transition semantic, the **cadbiom** software provides a formal framework to help the modeling of biological systems such as cell signaling network. It allows investigating synchronization events in biological networks. In 2016, the tool was integrated in a docker container in order to facilitate its use on any platform [49] [docker container] [web server].

- **The cadbiom graphical interface** is useful to build and study moderate size models. It provides means for model exploration, simulation and checking. For large-scale models, the graphical interface allows to focus on specific nodes of interest.
- **The cadbiom API** allows to load a model (including large-scale ones), perform static analysis (exploration, frontier computation, statistics, and dependence graph computation) and check temporal properties on a finite horizon in the future or in the past.
- **Exploring large-scale knowledge repositories** A main feature of cadbiom is that automatic translation of the large-scale PID repository (about 10,000 curated interactions) have been automatically translated into the cadbiom formalism. Therefore, the API allows for computing the upstream regulators of any set of genes based on this large-scale repository.

6.4.5. Protomata - Expressive pattern discovery on protein sequences

Protomata is a machine learning suite for the inference of *automata* characterizing (functional) families of proteins from available sequences. Based on a new kind of alignment said partial and local, it learns precise characterizations of the families – beyond the scope of classical sequence patterns such as PSSM, Profile HMM, or Prosite Patterns – allowing to predict new family members with a high specificity.

[1] http://www.cellnopt.org/

Protomata gives access to the three main modules as stand-alone programs, which are also integrated in a single workflow *protomata-learner*:

- **Paloma** builds partial local multiple alignments;
- **Protobuild** infers automata from these alignements;
- **Protomatch and Protoalign** scan, parse and align new sequences based on the automata infered previously. This module was improved in 2016 by embedding new options to score the sequences with respect to all accepting paths (Forward score) in addition to the scoring module based on the best path (Viterbi score). More generally, we have worked on the efficiency of the automata's weighting scheme based on the state-of-the-art schemes used for profile HMMs.

The suite is completed by many tools to handle or visualize data and can be used online via a [web interface].

6.4.6. Logol - Complex pattern modelling and matching

The **Logol** toolbox is a swiss-army-knife for pattern matching on DNA/RNA/Protein sequences, using a high-level grammatical formalism to permit a large expressivity for patterns [54]. A Logol pattern consists in a complex combination of motifs (such as degenerated strings) and structures (such as imperfect stem-loop ou repeats). Compared to other specialized pattern matching tools, some of the Logol key features are the possibilities to divide a pattern description into several sub-patterns, to enable the use of ambiguous models or to permit the inclusion of negative conditions in a pattern definition. Possible fields of application are the detection of mutated binding sites [32] or stem-loop identification (e.g. in CRISPR [2] [10]) [web page].

- The **Graphical designer** allows a user to iteratively build a complex pattern based on basic graphical patterns. The associated grammar file is an export of the graphical designer. In 2015, the efficiency of the tool was improved by slight evolutions of the underlying grammar.
- The **LogolMatch** parser takes as input a biological (nucleic or amino acid) sequence and a grammar file (i.e. a pattern). It combines a grammar analyzer, a sequence analyzer and a prolog Library. It returns a file containing all the occurrences of the pattern in the sequence with their parsing details.
- Full genome analysis, and connection to biological databases have been made available recently.

7. New Results

7.1. Data integration and pre-processing with semantic-based technologies

Participants: Meziane Aite, Marie Chevallier, Olivier Dameron, Aurélie Evrard, Clémence Frioux, Xavier Garnier, Jeanne Got, François Moreews, Yann Rivault, Anne Siegel, Pierre Vignet, Denis Tagu, Camille Trottier.

Integration and query of biological datasets with Semantic Web technologies. The purpose of this work is to obtain quick answers to biological questions demanding currently hours of manual search in several spreadsheet results files. We introduce an integration and interogation framework using an RDF model and the SPARQL query language. It allows biologists to transparently integrate and query their data without any a priori proficiency about RDF and SPARQL. [*O. Dameron, A. Evrard, X. Garnier*] [37], [45]

[2] http://crispi.genouest.org/

Handling the heterogeneity of genomic and metabolic networks data within flexible workflows with the PADMet toolbox A main challenge of the era of fast and massive genome sequencing is to transform sequences into biological knowledge. The high diversity of input files and tools required to run any metabolic networks reconstruction protocol represents an important drawback: it appears very difficult to ensure that input files agree among them. Such a heterogeneity produces loss of information during the use of the protocols and generates uncertainty in the final metabolic model. Here we introduce the PADMet-toolbox which allows conciliating genomic and metabolic network information. The toolbox centralizes all this information in a new graph-based format: PADMet (PortAble Database for Metabolism) and provides methods to import, update and export information. For the sake of illustration, the toolbox was used to create a workflow, named AuReMe, aiming to produce high-quality genome-scale metabolic networks and eventually input files to feed most platforms involved in metabolic network analyses. We applied this approach to two exotic organisms and our results evidenced the need of combining approaches and reconciling information to obtain a functional metabolic network to produce biomass. [*M. Chevallier, M. Aite, C. Frioux, J. Got, A. Siegel, C. Trottier, P. Vignet*] [34]

PEPS: a platform for supporting studies in pharmaco-epidemiology using medico-administrative databases We showed that Semantic Web technologies are technically adapted for representing patients' data from medico-administrative databases as RDF and querying them using SPARQL. We also demonstrated that this approach is relevant as it supports the combination of patients' data with hierarchical knowledge in order to address the problem of reconciling precise patients data with more general query criteria. [*O. Dameron, Y. Rivault*] [33], [31], [30]

Telemedicine : ontology-based reasoning and data integration We have developed a system based on a formal ontology that integrates the alert information and the patient data extracted from the electronic health record in order to better classify the importance of alerts. A pilot study was conducted on atrial fibrillation alerts. The results suggest that this approach has the potential to significantly reduce the alert burden in telecardiology. The methods may be extended to other types of connected devices. We also worked on a telemedicine application for monitoring patients with chronic diseases. We proposed an architecture supporting data exchange in the context of multiple chronic diseases [*O. Dameron*] [26], [25], [18]

7.2. Data and knowledge integration based on combinatorial optimization

Participants: Marie Chevallier, Damien Eveillard, Jeanne Got, Julie Laniau, François Moreews, Jacques Nicolas, Anne Siegel.

Packing graphs with ASP for landscape simulation This study is part of a more general research track on graph compression, a fundamental issue for the analysis of biological networks that we address with Answer Set Programming (ASP)modelling. The general issue is to cover a given graph by a set of subgraphs. The IJCAI paper describes an application to crop allocation for generating realistic landscapes. The aim is to cover optimally a bare landscape, represented by its plot graph, with spatial patterns describing local arrangements of crops. This problem belongs to the hard class of graph packing problems. The approach provides a compact solution to the basic problem and at the same time allows extensions such as a flexible integration of expert knowledge. Particular attention is paid to the treatment of symmetries, especially due to sub-graph isomorphism issues. Experiments were conducted on a database of simulated and real landscapes. Currently, our program can process graphs of medium size, a size that enables studies on real agricultural practices. [*J. Nicolas*] [29]

Deciphering transcriptional regulations coordinating the response to environmental changes We introduce a method that extracts from a transcriptional regulatory network determined from a set of predicted transcription factors (TF) and binding site (BS) a subnetwork explaining a given set of observed co-expressions, highlighting those TFs and BSs most likely involved in the co-regulation. The method solves an optimization problem on a graph to select confident paths within the given transcriptional regulatory network joining a putative common regulator with two co-expressed genes via regulatory cascades. It provides a useful modeling scheme for deciphering the regulatory mechanisms that underly the phenotypical response of an organism to

environmental challenges and can be used as a reliable tool for further research on genome scale transcriptional regulation studies. [*M. Chevallier, D. Eveillard, A. Siegel*] [13]

Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. Our software tool *shogen* [62] was used to decipher functional roles within a consortium of five mining bacteria through the integration of genomic and metabolic knowledge at genome scale. We first reconstructed a global metabolic network. Next, using a parsimony assumption, we deciphered sets of genes, called Sets from Genome Segments (SGS), that (i) are close on their respective genomes, (ii) take an active part in metabolic pathways and (iii) whose associated metabolic reactions are also closely connected within metabolic networks. The use of SGS (shogen) pinpoints a functional compartmentalization among the investigated species and exhibits putative bacterial interactions necessary for promoting these pathways. [*M. Chevallier, D. Eveillard, A. Siegel*] [15]

Molecular alterations induced by a high-fat high-fiber diet in porcine adipose tissues: variations according to the anatomical fat location Our methods based on the integration of metabolic and regulatory regulations [61] were combined to statistical approaches and applied to the understanding of fatty acid metabolism in porcs and chicken. The analyses evidenced that a high-fat high-fiber diet depressed glucose and lipid anabolic molecular pathways, thus counteracting adipose tissue expansion. Interaction effects between dietary intake of fiber and lipids on gene expression may modulate innate immunity and inflammation, a response which is of interest with regard to chronic inflammation and its adverse effects on health and performance. [*F. Moreews, A. Siegel*] [20]

7.3. Systems biology

Participants: Jérémie Bourdon, Jean Coquet, Victorien Delannée, Jacques Nicolas, Anne Siegel, Nathalie Théret, Pierre Vignet.

Representation of symbolic dynamical systems generated by a substitution. Iterated morphisms are combinatorial processes which are related to several classes of dynamical systems appearing in several fields of computer sciences and mathematics: numeration, ergodic theory, discrete geometry. They may be associated to fractal sets called "Rauzy fractals" whose topological properties are linked to the properties of the underlying dynamical system. We have introduced a generic algorithm framework to check such topological properties within a complete family of iterated morphism. This makes efficient the verification of conjectures on several families of substitutions related to multi-dimensional continued fraction algorithms. [*A. Siegel*] [14]

Identification of logical models for signaling pathways. Logical models of signaling pathways are a promising way of building effective in silico functional models of a cell. The automated learning of Boolean logic models describing signaling pathways can be achieved by training to phosphoproteomics data. In our work, combinatorial optimization methods based on recent logic programming paradigm allow to enumerate, and discriminate the family of logical models explaining data. Together, these approaches enable a robust understanding of the system response. The results are implemented in the *caspo* software. The main weakness of ASP-based learning algorithm is that they focus on the comparison of two time-points and assumes that the system has reached an early steady state. We have generalized such a learning procedure in order to discriminate Boolean networks according to their transient dynamics. To that goal, we exhibit a necessary condition that must be satisfied by a Boolean network dynamics to be consistent with a discretized time series trace. [*A. Siegel*] [23], [28]

Model of the Delayed Translation of Cyclin B Maternal mRNA After Sea Urchin Fertilization. An extended model of the numerical model introduced in [74] was developed to have a better understanding of the role of cyclin B in protein synthesis within minutes after fertilization of sea urchin eggs. The model confirms that regulation of cyclin B biosynthesis is an example of a select protein whose translation is controlled by pathways that are distinct from housekeeping proteins, even though both involve the same cap-dependent initia- tion pathway. Therefore, this model should help provide insight to the signaling utilized for the biosynthesis of cyclin B and other select proteins. [*J. Bourdon, A. Siegel*] [24]

Deciphering pathways involved in TGF-β signalling network. TGF-β is a multifunctional cytokine that regulates mammalian development, differentiation, and homeostasis. As a growth inhibitor of epithelial, endothelial, and hematopoietic cells, TGF-β is a potent anticancer agent in normal tissue. At the opposite TGF-β acts as a promoter of tumor by inducing the hallmarks of the cancer. Consequently targeting the deleterious effects of TGF-β without affecting its physiological role is the common goal of therapeutic strategies. While several strategies based on blocking TGF-β antibodies or small inhibitors of TGF-β receptors have been investigated, they did not take into account the impact of the (extracellular matrix) ECM remodeling that regulates TGF-β bioavailability and the complexity of TGF-β-dependent signaling pathways which regulate both physiological and pathological processes depending on context. In accordance with this, we recently demonstrated the benefical anti-tumor effect of the interplay between TGF-β signaling and the CD103 integrin pathway. At the opposite we have previously demonstrated that the disintegrin ADAMTS1 promotes TGF-β activation in chronic liver disease and we recently characterized interaction with inhibitor peptide to block such effects, using in silico approach. Importantly, we need to take into account a system-wide view and develop predictive models for therapeutic benefit. In that context we demonstrated that the ratio of TGFBR2 to TGFBR1 receptors concentrations can be used to discriminate between metastable regimes of TGF-β signaling model and predic the tumor cell aggressiveness [*N. Théret*][27], [16], [21].

7.4. Sequence and structure annotation

Participants: Aymeric Antoine-Lorquin, Catherine Belleannée, François Coste, Jacques Nicolas.

Detection of mutated primers on metagenomics sequences to detect more species. In targeted metagenomics, an initial task is the detection in each sequence of the primers used for amplifying the targeted region. The selected sequences are then trimmed and clustered in order to inventory species present in the sample. Common pratices consist in retaining only the sequences with perfect primers (i.e. non-mutated by sequencing error). In the context of a study characterizing the biodiversity of tropical soils in unicellular eukaryotes, we have implemented the search for mutated primers, using the grammatical pattern matching tool Logol, and shown that retrieving sequences with mutated primers has a significant impact on targeted metagenomics results, as it makes possible to detect more species (7% additional OTUs in our study) [*A. Antoine-Lorquin, C. Belleannée*] [32], [11].

VIRALpro: a tool to identify viral capsid and tail sequences. Not only sequence data continues to outpace annotation information, but the problem is further exacerbated when organisms are underrepresented in the annotation databases. This is the case with non human-pathogenic viruses which occur frequently in metagenomic projects. Thus there is a need for tools capable of detecting and classifying viral sequences. Based on machine learning techniques, we have proposed a new effective tool for identifying capsid and tail protein sequences, which are the cornerstones toward viral sequence annotation and viral genome classification. The software and corresponding web server are publicly available as part of the SCRATCH suite. [*F. Coste, C. Galiez*] [19]

Learning substitutable context-free grammars to model protein families. In the first experiments on learning substitutable context-free grammars to model protein families, an identified bottleneck for larger scale experimentation was parsing time. We have implemented a new parsing strategy enabling to handle efficiently the ambiguity of 'gap loops', enabling a factor 20 speedup in practice. We have also begun to investigate the inference of more expressive classes, said contextually substitutable, and have proposed a refined graph approach to learn smaller contextually substitutable grammars from smaller training samples in the framework that we have initiated with ReGLiS. [*F. Coste*] [43], [35]

How to measure the topological quality of protein grammars? To assess the quality of grammars modelling protein families, one is interested in their performances to predict new members of the families, classically measured on the basis of recall and precision in the machine learning framework, but also by their modelling power, which is more difficult to evaluate. We propose here to address this later point by measuring the consistency of grammar's parse trees with 3D structures of proteins, when they are available, by the introduction of a set of measures based on respective internal distances. [*F. Coste*] [36]

Tutorial chapter: Learning the Language of Biological Sequences. Learning the language of biological sequences is an appealing challenge for the grammatical inference research field. While some first successes have already been recorded, such as the inference of profile hidden Markov models or stochastic context-free grammars which are now part of the classical bioinformatics toolbox, it is still a source of open and nice inspirational problems for grammatical inference, enabling us to confront our ideas to real fundamental applications. As an introduction to this field, we survey here the main ideas and concepts behind the approaches developed in pattern/motif discovery and grammatical inference to characterize successfully the biological sequences with their specificities. [*F. Coste*] [40]

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Regional initiative: the Ecosyst project

Participants: Damien Eveillard, Marie Chevallier, Clémence Frioux, Anne Siegel, Camille Trottier.

EcoSyst is a Biogenouest inter-regional federating project (Brittany â Pays de la Loire) aiming at the emergence of Systems Ecology at the level of western regions. Drawing on the strengths and skills involved, EcoSyst targets the incubation of new ideas and new projects at disciplinary interfaces. Thanks to this community project, we want to develop the skills of Ecology, Environment, Modeling, Bioinformatics and Systems Biology and their application to organisms and ecosystems of interest in agronomy, sea and health. EcoSyst includes also the identification of the major issues and concerns, the fundamental and essential methods and the very real needs of the community (training, tools, ...); this in order to consider the construction of a community platform (or an offer of service within an existing platform) on complex systems modeling, meeting expectations of the community as fully as possible.

8.1.2. Regional partnership with computer science laboratories in Nantes

Participants: Anne Siegel, Jérémie Bourdon, Damien Eveillard, François Coste, Jacques Nicolas.

Methodologies are developed in close collaboration with university of Nantes (LINA) and Ecole centrale Nantes (IRCCyN). This is acted through the Biotempo and Idealg ANR projects and co-development of common software toolboxes within the Renabi-GO platform support. C. Trottier is a co-supervised bioianalysis and software development engineer within the Idealg project. M. Chevallier is a co-supervised development and animation engineer within the regional initiative "Ecosyst". In addition, the former Ph-D student V. Picard and the ongoing Ph-D student J. Laniau are also co-supervised with members of the LINA laboratory.

8.1.3. Regional partnership in Marine Biology

Participants: Catherine Belleannée, Jérémie Bourdon, Jean Coquet, François Coste, Damien Eveillard, Olivier Dameron, Clémence Frioux, Jeanne Got, Julie Laniau, Jacques Nicolas, Camille Trottier, Anne Siegel.

A strong application domain of the Dyliss project is marine Biology. This application domain is co-developed with the station biologique de Roscoff and their three UMR and involves several contracts. The IDEALG consortium is a long term project (10 years, ANR Investissement avenir) aiming at the development of macro-algae biotechnology. Among the research activities, we are particularly interested in the analysis and reconstruction of metabolism and the characterization of key enzymes. Other research contracts concern the modeling of the initiation of sea-urchin translation (former PEPS program Quantoursin, Ligue contre le cancer and ANR Biotempo), the analysis of extremophile archebacteria genomes and their PPI networks (former ANR MODULOME and PhD thesis of P.-F. Pluchon) and the identification of key actors implied in competition for light in the ocean (PELICAN ANR project). In addition, the team participates to a collaboration program (Inria Project Lab) with the Biocore and Ange teams, together with Ifremer-Nantes, focused on the understanding on micro-algae (Ph-D thesis of J. Laniau).

8.1.4. Regional partnership in agriculture and bio-medical domains

Participants: Aymeric Antoine-Lorquin, Catherine Belleannée, François Coste, Jean Coquet, Olivier Dameron, Victorien Delannée, Aurélie Evrard, François Moreews, Jacques Nicolas, Anne Siegel, Nathalie Théret, Denis Tagu, Pierre Vignet.

We have a strong and long term collaboration with biologists of INRA in Rennes : PEGASE and IGEPP units. F. Morrews is a permanent engineer from PEGASE center hosted in the team to develop methods for integrative biology applied to species of interest in agriculture. D. Tagu is a research director at INRA who spends 20% of his time in the team to develop collaborative projects. This partnership is supported by the co-supervision of one post-doctoral student and the co-supervision of several PhD students. The Ph-D thesis of V. Wucher was supported by collaborations with the IGEPP laboratory. The former post-doc of Ch. Bettembourg strengthened these collaborations. This collaboration was also reinforced by collaboration within ANR contracts (MirNadapt, FatInteger). Lately, A. Evrard joined the team at mid-part of her time in collaboration with Agrocampus Ouest and INRA to apply the semantic web to technologies developed within the mirnAdapt framework to new agriculture applications (Brassicae).

We also have a strong and long term collaboration in the bio-medical domain, namely with the IRSET laboratory at Univ. Rennes 1/Irset. N. Théret, research director at INSERM, is hosted in the team to strenghen our collaborative projects. Our collaborations are acted by the co-supervised Ph-D theses of V. Delannée (Metagenotox project, funded by Anses) and J. Coquet. This partnership was reinforced in the former years by the ANR contract Biotempo ended at the end of 2014. In 2015, the project of combining semantic web technologies and bi-clustering classification based on formal concept analysis was applied to systems biology within the PEPS CONFOCAL project. This scientific project has been recently pushed forward in the recent TGFSYSBio project funded by Plan Cancer on the modelling of the microenvironment of TGFbeta signaling network (P. Vignet has been recruited on this contract at the end of 2016).

8.2. National Initiatives

8.2.1. Long-term contracts

8.2.1.1. "Omics"-Line of the Chilean CIRIC-Inria Center

Participants: Meziane Aite, Jérémie Bourdon, François Coste, Marie Chevallier, Damien Eveillard, Clémence Frioux, Jacques Nicolas, Anne Siegel.

We have a cooperation with Univ. of Chile (MATHomics, A. Maass) on methods for the identification of biomarkers and software for biochip design, supported by a national Inria initiative. It aims at combining automatic reasoning on biological sequences and networks with probabilistic approaches to manage, explore and integrate large sets of heterogeneous omics data into networks of interactions allowing to produce biomarkers, with a main application to biomining bacteria. The program is co-funded by Inria and CORFO-chile from 2012 to 2022. In this context, IntegrativeBioChile is an Associate Team between Dyliss and the Laboratory of Bioinformatics and Mathematics of the Genome hosted at Univ. of Chile funded from 2011 to 2016.

8.2.1.2. ANR Idealg

Participants: Jérémie Bourdon, Marie Chevallier, François Coste, Damien Eveillard, Clémence Frioux, Jeanne Got, Jacques Nicolas, Anne Siegel.

IDEALG is one of the five laureates from the national call 2010 for Biotechnology and Bioresource and will run until 2020. It gathers 18 different partners from the academic field (CNRS, IFREMER, UEB, UBO, UBS, ENSCR, University of Nantes, INRA, AgroCampus), the industrial field (C-WEED, Bezhin Rosko, Aleor, France Haliotis, DuPont) as well as a technical center specialized in seaweeds (CEVA) in order to foster biotechnology applications within the seaweed field. It is organized in ten workpackages. We are participating in the tasks related to the establishment of a virtual platform for integrating omics studies on seaweed) and the integrative analysis of seaweed metabolism, in cooperation with SBR Roscoff. Major objectives are the building of brown algae metabolic maps, flux analysis and the selection extraction of important parameters for the production of targeted compounds. We will also contribute to the prediction of specific enzymes (sulfatases) [More details].

8.2.2. Programs funded by research institutions

8.2.2.1. PEPS PEPS: a platform for supporting studies in pharmaco-epidemiology using medico-administrative databases

Participants: Olivier Dameron, Yann Rivault.

As a partner of the PEPS platform, IRISA develops generic methods supporting efficient and semantically-rich queries for pharmaco-epidemiology studies on medico-administrative databases. The leader is Thomas Guyet (IRISA team Lacodam). We showed that Semantic Web technologies are technically suited for representing patients' data from medico-administrative databases as RDF and querying them using SPARQL. We also demonstrated that this approach is relevant as it supports the combination of patients' data with hierarchical knowledge in order to address the problem of reconciling precise patients data with more general query criteria [33], [31], [30]. This work is mostly conducted by Yann Rivault, whose PhD thesis is supervized by Olivier Dameron and Nolwenn LeMeur (Ecole des Hautes Etudes en Santé Publique).

8.2.2.2. Cancer Plan: TGFSYSBIO

Participants: Nathalie Théret, Jacques Nicolas, Olivier Dameron, Anne Siegel, Jean Coquet.

The TGFSYSBIO project aims to develop the first model of extracellular and intracellular TGF-beta system that might permit to analyze the behaviors of TGF-beta activity during the course of liver tumor progression and to identify new biomarkers and potential therapeutic targets. Based on collaboration with Jerome Feret from ENS, Paris,we will combine a rule-based model (Kappa language) to describe extracellular TGF-beta activation and large-scale state-transition based (Cadbiom formalism) model for TGF-beta-dependent intracellular signaling pathways. The multi-scale integrated model will be enriched with a large-scale analysis of liver tissues using shotgun proteomics to characterize protein networks from tumor microenvironment whose remodeling is responsible for extracellular activation of TGF-beta. The trajectories and upstream regulators of the final model will be analyzed with symbolic model checking techniques and abstract interpretation combined with causality analysis. Candidates will be classified with semantic-based approaches and symbolic bi-clustering technics. The project is funded by the national program "Plan Cancer - Systems biology" from 2015 to 2018.

8.2.2.3. ANR Samosa

Participants: Damien Eveillard, Jeanne Got, Anne Siegel.

Oceans are particularly affected by global change, which can cause e.g. increases in average sea temperature and in UV radiation fluxes onto ocean surface or a shrinkage of nutrient-rich areas. This raises the question of the capacity of marine photosynthetic microorganisms to cope with these environmental changes both at short term (physiological plasticity) and long term (e.g. gene alterations or acquisitions causing changes in fitness in a specific niche). Synechococcus cyanobacteria are among the most pertinent biological models to tackle this question, because of their ubiquity and wide abundance in the field, which allows them to be studied at all levels of organization from genes to the global ocean.

The SAMOSA project is funded by ANR from 2014 to 2018, coordinated by F. Gaczarek at the Station Biologique de Roscoff/UPMC/CNRS. The goal of the project is to develop a systems biology approach to characterize and model the main acclimation (i.e., physiological) and adaptation (i.e. evolutionary) mechanisms involved in the differential responses of Synechococcus clades/ecotypes to environmental fluctuations, with the goal to better predict their respective adaptability, and hence dynamics and distribution, in the context of global change. For this purpose, following intensive omics experimental protocol driven by our colleagues from − Station Biologique de Roscoff −, we aim at constructing a gene network model sufficiently flexible to allow the integration of transcriptomic and physiological data.

8.2.2.4. ADT Complex-biomarkers and ADT Proof of concept

Participants: Jeanne Got, Marie Chevallier, Meziane Aite, Anne Siegel.

This project started in Oct. 2014 and aims at designing a working environment based on workflows to assist molecular biologists to integrate large-scale omics data on non-classical species. The main goal of the workflows will be to facilitate the identification of set of regulators involved in the response of a species when challenged by an environmental stress. Applications target extremophile biotechnologies (biomining) and marine biology (micro-algae).

8.2.2.5. ANSES Mecagenotox

Participants: Victorien Delannée, Anne Siegel, Nathalie Théret.

The objective of Mecagenotox project is to characterize and model the human liver ability to bioactivate environmental contaminants during liver chronic diseases in order to assess individual susceptibility. Indeed, liver pathologies which result in the development of fibrosis are associated with a severe dysfunction of liver functions that may lead to increased susceptibility against contaminants. In this project funded by ANSES and coordinated by S. Langouet at IRSET/inserm (Univ. Rennes 1), we will combine cell biology approaches, biochemistry, biophysics, analytical chemistry and bioinformatics to 1) understand how the tension forces induced by the development of liver fibrosis alter the susceptibility of hepatocytes to certain genotoxic chemicals (especially Heterocyclic Aromatic Amines) and 2) model the behavior of xenobiotic metabolism during the liver fibrosis. Our main goal is to identify "sensitive" biomolecules in the network and to understand more comprehensively bioactivation of environmental contaminants involved in the onset of hepatocellular carcinoma.

8.2.2.6. PEPS CONFOCAL

Participants: Olivier Dameron, Jean Coquet, Nathalie Théret, Jacques Nicolas, Anne Siegel, Pierre Vignet.

PEPS CONFOCAL aims at developing new bioinformatics methods for analyzing heterogeneous *omics data and for filtering them according to domain knowledge. The current approaches are facing four main limitations: (1) classic biclustering methods do not support partial overlap of clusters, which is too restrictive considering some genes' pleiotropic nature, (2) they assume that the items to analyze (the genes, the molecules, the signaling pathways...) are independent, (3) they tend to generate numerous clusters leaving to the experts the task of identifying the relevant ones, and (4) they are sensitive to noisy or incomplete data. We investigate the extension of Formal Concept Analysis (FCA) with symbolic knowledge from ontologies in order to process large and complex sets of associations between genes, signaling pathways and the molecules involved in these pathways. Future applications cover the discrete model analysis in molecular biology. CONFOCAL initiated a collaboration with Amedeo Napoli (LORIA Nancy) and Elisabeth Remy (Mathematics Institute Luminy, "Mathematical Methods for Genomics" team).

8.3. European Initiatives

8.3.1. Collaborations with Major European Organizations

Partner: Aachen university (Germany)

Title: Modeling the logical response of a signalling network with constraints-programming.

Partner: Potsdam university (Germany)

Title: Constraint-based programming for the modeling and study of biological networks.

8.4. International Initiatives

8.4.1. Inria International Labs

The Dyliss team is strongly involved in the Inria CIRIC center, and the research line "Omics integrative center". The associated team "IntegrativeBioChile", the post-doc of S. Thiele (2012) and the co-supervision of A. Aravena (2010-2013) contributed to reinforce the complementarity of both Chilean and French teams. In 2013, a workshop was organized in Chile to develop new French-Chilean collaborations within the framework of the CIRIC center. In 2014, Marie Chevallier and Meziane Aite joined the team as engineers to improve softwares resulting from collaborations. Maria-Paz Cortes visited the team during 6 monthes in the framework of her ph-D thesis.

Inria Chile
Associate Team involved in the International Lab:

8.4.1.1. BIOINTEGRATIVECHILE

Title: Integrative Biology in Extreme Environments

International Partner (Institution - Laboratory - Researcher):

Universidad de Chile (Chile) - Center for Mathematical Modeling (CMM) - Maass Alejandro

Start year: 2014

See also: http://www.irisa.fr/dyliss/public/EA/index.html

The project is in the area of bioinformatics, with a special focus on bacteria living in extreme environments, more precisely on microorganisms involved in bio-remediation or bio-production processes. We are particularly interested in bioprocesses such as copper extraction, salmon lethality, metal-resistance, all having an economical interest in Chile. Since the last decade, huge databases of microbial genomic sequences, together with multi-scale and large-scale cellular observations (genomics, transcriptomics, proteomics, metabolomics) have been produced. Each one can be viewed as a different scale of a biological process, either in time or space, but ultimately are related through networks of biological interactions that control the behavior of the system The reconstruction, analysis and modeling of such networks using all levels of information are biologically, mathematically and computationally challenging. Applied on microorganisms living in extreme environments, this question is even more challenging since relatively few knowledge is publicly available on the species, requiring to develop methods which are robust to uncertainty. We are developing methods to integrate and manage heterogeneous omics and uncertain data. This in the purpose of extracting suitable biomarkers from this multi-level information. This question will be addressed by coupling probabilistic and static dynamical systems methods with recent and efficient paradigms of constraint programming (Answer Set Programming).

8.5. International Research Visitors

8.5.1. Visits of International Scientists

- **Argentina**. Fundation Leloir, Buenos Aeres [S. Videla]
- **Chile**. Centro de Modelimiento Matematico, Santiago [A. Maass, N. Loiraâ , M. Latorre, M.-P. Cortes]
- **Niger**. University of Maradi [O. Abdou-Arbi]
- **Germany**. Max Planck Institute for Biophysical Chemistry [C. Galiez]

8.5.2. Research stays abroad

- **Germany**. University of Kaiserslautern [A. Antoine-Lorquin, 2 months]
- **Germany**. University of Potsdam [C. Frioux, 2 months]
- **Japan**. National Institute of Informatics in Tokyo [J.Coquet, 3 months]

8.5.3. Visits to International Teams

- **Chile**. Centro de Modelimiento Matematico, Santiago de Chile [J. Bourdon, M. Aite, F. Coste, A. Siegel]
- **Germany**. Frei Berlin University [A. Siegel]
- **Poland**. Wroclaw University of Science and Technology [F. Coste]
- **Netherland**. Utrecht University [F. Coste]

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Selection

9.1.1.1. Member of the Conference Program Committees

- JOBIM (2016): French conference of Bioinformatics [A. Siegel]
- BBCC (2015): Bioinformatica e Biologia Computazionale in Campania [O. Dameron]
- IC 2016 Atelier IA et santé, Symposium sur l'Ingénierie des Connaissances [O. Dameron]
- JFO 2016: Journées Francophones sur les ontologies [O. Dameron]
- SASB 2016: The six international workshop on static analysis and system biology [N.Théret]
- ICGI 2016: The 13th International Conference on Grammatical Inference [F. Coste]

9.1.1.2. Review

- ISMB 2016, Biotechno 2016 [A. Siegel]

9.1.2. Journal

9.1.2.1. Member of the Editorial Boards

- Academic editor: Plos One [J. Bourdon]

9.1.2.2. Reviewer - Reviewing Activities

- Journal of Mathematical Biology. Bioinformatics. Theorie des Sciences Informatiques. [A. Siegel]
- Briefings in Bioinformatics, Journal of Biomedical Informatics, Journal of Biomedical Semantics. [O. Dameron]
- Molecular Cancer, Oncotarget,Hepatology, Int J cancer, Bioinformatics and Biology Insight, Chembioint, Cell death and Disease, Plos One [N. Théret]
- IEEE BIBM 2016: IEEE International Conference on Bioinformatics and Biomedicine [F. Coste]

9.1.3. Invited Talks

- M. Aite *User-control metabolic network reconstruction within flexible workflows with the PADMet Toolbox*, INRA Food Working Group annual assembly, Paris (Jul. 2016)
- F. Coste *Grammatical inference of protein languages*, Seminar of Department of Biomedical Engineering, Wroclaw University of Technology, Polland (May 2016)
- F. Coste *Partial multiple sequence alignments to model protein families*, Theoretical Biology and Bioinformatics Group, Utrecht Univ., Netherland (Sep. 1016)
- F. Coste *Modelling protein families with Protomata-Learner*, University of Chile (Nov. 2016)
- C. Trottier *The PADMET-Toolbox and AuReMe workflow: application to the genome-scale metabolic network reconstruction of algae*, IDEALG project Annual General Meeting, Lorient (Nov. 2016)
- C. Frioux *Metabolic network gap-filling: parsimonious combinatorial methods to approach biological reality*, INRA Food Working Group annual assembly, Paris (Jul. 2016)
- C. Frioux *Answer Set Programming for bioinformatics and metabolic networks*, University of Potsdam Knowledge Processing and Information System group weekly seminar, Potsdam, Germany (Oct. 2016)
- A. Siegel *Cancer biology in the Dyliss Group*, Inria (Mar. 2016)
- A. Siegel *Identification of logical models for signaling pathways: towards a systems biology loop*, EMCTS, Nottingham, UK (Jul. 2016) [39]
- A. Siegel *An introduction to metabolic networks modelling*, INRA (Oct. 2016)

- A. Siegel *A prospective about the construction of bioId chip based on multi-scale integrative methods*, University of Chile (Nov. 2016)
- A. Siegel *Combinatorial problems related to the reconstruction of genome-scale metabolic networks*, University of Lille, Workshop of the BIOSS working group on metabolism (Nov. 2016)
- N.Théret *Computational modeling to identify biomarkers and targets*, DHU2020, Fibrosis and remodelling: from common pathways to personalized targets, Autumn School (Oct. 2016)

9.1.4. Leadership within the Scientific Community

- Member of the steering committee of the International Conference on Grammatical Inference [F. Coste].
- The team was involved in the foundation of a national working group on the symbolic study of dynamical systems named bioss [web access]. The group gathers more 100 scientists, from computer science to biology. Three meetings were organized this year. The group is supported by two French National Research Networks: bioinformatics (GDR BIM : bioinformatique moléculaire) and informatics-mathematics (GDR IM : Informatique Mathématique). It gathered twice in 2016: for a general meeting in Lyon (Jul. 2016) and for a workshop focused on computational and methological insights about metabolic network in Lille (Nov. 2016) [A. Siegel]

9.1.5. Scientific Expertise

- Member (nominated) CNU section 65 [O. Dameron]
- Scientific Advisory Board of GDR BIM " Molecular Bioinformatics" [J. Nicolas].
- Inria National evaluation board [A. Siegel]
- Member of the Operational Legal and Ethical Risk Assessment Committee (COERLE) at Inria [J. Nicolas].
- Recruitment committees: Professor (UMPC, Paris) [A. Siegel, N Théret], Associate professor (Nancy) [O. Dameron), Engineer (INRA) [A. Siegel], Inria senior researcher (National committee) [A. Siegel].
- Member of the IRISA laboratory council [F. Coste].
- Member of the Inria Rennes center council [A. Siegel].
- Scientific Advisory Board of Biogenouest [J. Bourdon, N Théret].
- Member of SCAS (Service Commun d'Action Sociale) of Univ. Rennes 1 [C. Belleannée].
- Member of CUMIR (Commission des Utilisateurs des Moyens Informatiques, Inria Rennes) [F. Coste].
- Expertise for Prix Victor et Erminia MESCLE [N. Théret]
- Expertise for Ligue Contre le Cancer, InterRégion Rhône-Alpes-Auvergne-Drome-Saone et Loire [N. Théret]
- Member of the board of directors of the French Society for biology of the extracellular matrix [N. Théret]

9.2. Teaching - Supervision - Juries

9.2.1. Teaching track responsibilities

- Coordination of the doctoral school "Life, Agronomy and Health" of University of Rennes 1 [N. Théret]
- Coordination of the master degree "Bioinformatics and genomics", Univ. Rennes1 [O. Dameron]
- Coordination of the track "From Data to Knowledge: Machine Learning, Modeling and Indexing Multimedia Contents and Symbolic Data", Master in Computer Science, University of Rennes 1, France [F. Coste].

9.2.2. Course responsibilities

- "Bioinformatique expérimentale", Master 1 in computer science, Univ. Rennes1 & ENS [O. Dameron]
- "Bases de mathématiques et probablité" and "Méthodes en informatique", Master1 in public health, Univ. Rennes 1 [O. Dameron]
- "Représentation des connaissances biomédicales", Master 2 in public health, Univ. Rennes 1 [O. Dameron]
- "Principes de programmation et d'algorithmique", Master 1 in bioinformatics, Univ. Rennes 1 [O. Dameron]
- "Gestion de projets informatiques", Master 1 in bioinformatics, Univ. Rennes 1 [O. Dameron]
- "Standardisation des connaissances et bio-ontologies", Master 2 in bioinformatics, Univ. Rennes 1 [O. Dameron]
- "e-Santé et réseaux hospitaliers", last year in engineering school ESIR, Univ. Rennes 1, [O. Dameron]
- "Equilibre Dynamique de la communication Cellulaire" Master 2 in Sciences cellulaire et Moléculaire du Vivant, Univ. Rennes 1 [N. Theret]

9.2.3. Teaching

Licence: C. Belleannée, Langages formels, 22h, L3 informatique, Univ. Rennes1, France.

Licence: C. Belleannée, Traitement de textes et données tabulées, 40h, L1 informatique, Univ. Rennes1, France.

Licence: C. Belleannée, Algorithmique et Programmation Fonctionnelle, 60, L1 informatique, Univ. Rennes1, France.

Licence: J. Coquet, Algorithmique et Programmation Fonctionnelle, 40h, L1 informatique, Rennes1, France.

Licence: O. Dameron, Biostatistiques, 12h, PACES, Univ. Rennes 1, France.

Licence: V. Delannée, Bureautique, 36h, DFGSP2, Univ. Rennes 1, France.

Licence: C. Frioux, Bureautique, 12h, L1 informatique, Rennes1, France.

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

Master: C. Belleannée, Préférences, Logique et Contraintes, 40h M1 informatique,Univ. Rennes1, France

Master: F. Coste, Apprentissage Supervisé, 10h, M2 Informatique, Univ. Rennes 1, France

Master: F. Coste, Données Séquentielles Symboliques, 10h, M2 Informatique, Univ. Rennes 1, France

Master: O. Dameron, Bases de mathématiques et probablité, 30h, Master1 in public health, Univ. Rennes 1, France.

Master: O. Dameron, Méthodes en informatique, 50h, Master1 in public health, Univ. Rennes 1, France.

Master: O. Dameron, Bioinformatique expérimentale, 10h, M1 informatique, Univ. Rennes 1 and ENS Rennes, France.

Master: O. Dameron, Principes de programmation et algorithmique, 50h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, Gestion de projets informatiques, 23h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, Standardisation des connaissances et bio-ontologies, 24h, M2 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, Représentation des connaissances biomédicales, 20h, M2 bioinformatique et génomique, Univ. Rennes 1, France.

Master: A. Siegel, Integrative and Systems biology, 20h, M2, Univ. Rennes 1, France

Master: N. Théret, Extracellular matrix remodeling and Signaling, 3H, Univ. Rennes 1, France

Master: N. Théret, Extracellular matrix remodeling and Signaling, 3H, Univ. Cergy Pontoise, France

Doctorat: A. Siegel, Modelling the integration of heterogeneous knowledge with Answer Set Programming, 4h, Ecole de printemps, Porquerolles, France

9.2.4. Supervision

HDR: Olivier Dameron *Ontology-based methods for analyzing life science data* [12]

PhD : Aymeric Antoine-Lorquin, *TITRE*, started in Oct. 2013, supervised by C. Belleannée, defended on the 1st of December 2016 [11]

PhD in progress : Lucas Bourneuf, *Justifiable graph decomposition to assist biological network understanding*, started in Oct. 2016, supervised by J. Nicolas.

PhD in progress : Jean Coquet, *Semantic-based reasoning for biological pathways analysis*, started in Oct. 2014, supervised by O. Dameron and N. Théret.

PhD in progress : Victorien Delannée, *Optimisation à différentes échelles pour étudier la variabilité de la toxicité de contaminants alimentaires*, started in Oct. 2014, supervised by A. Siegel and N. Théret.

PhD in progress : Clémence Frioux, *Using preferences in Answer Set Programming to decipher interactions within the species of an ecosystem at the genomic scale*, started in Oct. 2015, supervised by A. Siegel.

PhD in progress : Julie Laniau, *Méthodes d'optimisation combinatoire pour reconstruire et analyser les systèmes métaboliques de microalgues*, started in Oct. 2013, supervised by A. Siegel and D. Eveillard.

PhD in progress : Yann Rivault, *Analyse de parcours de soins à partir de bases de données médico-administratives en utilisant des outils du Web Sémantique : identification de complications et de leurs déterminants suite à la pose chirurgicale de dispositif médical implantable en ambulatoire* , started in Oct. 2015, supervised by O. Dameron and N. Lemeur.

9.2.5. Juries

- *Member of Ph-D thesis juries*. M. Morterol, Univ. Paris Sud [A. Siegel, reviewer]. A. Rougny, Univ. Paris Sud [A. Siegel, reviewer]. P. Traynard, ENS Paris [A. Siegel, jury member]. A Lamora, Univ. Nantes [N. Théret, reviewer]. L. Alcaraz, Univ. Lyon [N. Théret, reviewer], F Courivaud, UMPC [N. Théret, reviewer]. P Hascoet, Univ. Rennes1 [N. Théret, president]
- *Member of habilitation thesis juries*. O. Dameron, Univ. Rennes 1 [A. Siegel, jury member], A. Chateau, Univ. Montpellier [A. Siegel, reviewer]. M. Elati, Univ. Evry [A. Siegel, reviewer]. L. Levy, Univ Paris-Diderot [N. Théret, reviewer]. C. Le Goff, Univ Paris Descartes [N. Théret, reviewer].
- *Member of medical thesis jury*. P. Hamon, Rennes [O. Dameron, jury member].

9.2.6. Internships

- Internship, from Jan until Jun 2016. Supervised by A. Siegel. Student: Mael Conan. Subject: Reconstruction of the metabolic map of E. Synecchococcus.
- Internship, from Jan until Jun 2016. Supervised by M. Chevallier and A. Siegel. Student: Pierre Vignet. Subject: Development of a web interface for the aided-curation of metabolic network identifiers.

- Internship, from Jun until Jul 2016. Supervised by O. Dameron and A. Siegel. Student: David Saulpic. Subject: Using formal concept analysis to classify the attractors of perturbated boolean networks.
- Internship, from Jan. until Jun 2016. Supervised by J. Nicolas. Student: Lucas Bourneuf. Subject: Model reduction with power graph algorithms.
- Internship, from Feb. until Jun 2016. Supervised by F. Coste. Student: Mikael Demirdelen. Subject: Fast parser for biological sequences and a new algorithm for the inference of substitutable languages.
- Internship, from Jun until Jul 2016. Supervised by O. Dameron. Student: Arnaud Belcour. Subject: Intégration de données biologiques en RDF pour l'analyse de réseaux de régulation.
- Internship, from Jun until Jul 2016. Supervised by O. Dameron. Student: Mael Kerbiriou. Subject: Création et analyse d'un réseau de régulation génique en RDF : application au puceron.
- Internship, from Jun until Jul 2016. Supervised by A. Evrard. Student: Xavier Garnier. Subject: Mise à jour et développement d'AskOmics, outil d'intégration et d'interrogation de données biologiques.
- Internship, from Jan. until Jun. 2016. Supervised by C. Belleannée. Student: Nathan Alary. Subject: Données génomiques et données ChIP-Seq au service de la prédiction de sites de fixation d'un facteur de transcription. Application au facteur LXRalpha.
- Internship, from Mar until Aug 2016. Supervised by J. Got. Student: Sanae El Mhijar, Subject: Analyse et vérifications du réseau métabolique de *Tisochrysis lutea*.
- Internship, from May until Jul 2016. Supervised by F. Coste. Student: RemySun, Subject: Learning Deep Latent Features of Proteins.
- Internship, from Jun. until Jul. 2016. Supervised by F. Morreews. Student: Vivien Le Breton Subject: RDF et SPARQL pour l'integration de réseaux métaboliques et génétiques de référence.
- Internship, from Jan. until Jun. 2016. Supervised by J. Nicolas. Student: Guillaume Lebreton Subject: Metabolic pathway reconstruction on metagenomes, application to the development of a bacterial consortium for fermented products.
- Internship, from Apr. until Jun. 2016. Supervised by J. Nicolas. Student: Marie Salmon Subject: Analyse par concepts formels de données génomiques sur le mélanome du chien.

9.3. Popularization

- *Organization of Sciences en Cour[t]s*. Since 2007, Sciences en Cour[t]s is a project of Nicomaque organization, the association of PhD and PhD students of Brittany. It is a popularization Festival where PhD students explain their thesis via short films of 5min. The goal is to present their scientific researches to the general public. Every year, PhD students of Inria/IRISA join the organization or make movies.[*J. Coquet (coordinator of the festival), V. Delannée (president of Nicomaque), A. Antoine-Lorquin (organizer of the festival)*] [more info].
- *Production of Sciences en Cour[t]s film*. "Une petite histoire de symbiose(s)" .[*C. Frioux*] [more info].
- *Bioinfo-fr.net* Bioinfo-fr.net is a french web site where researchers, engineers and students talks about bioinformatics. We have written or contributed to 3 articles for this web site on diverse subjects: "Remise des diplômes du master BIG (Rennes)", "Les dev' jam c'est bon pour vous !", "Ecrire son parseur à la main: chroniques d'une mauvaise bonne idée". [*L Bourneuf, O. Dameron*]. [more info].

10. Bibliography

Major publications by the team in recent years

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

[2] J. BOURDON, D. EVEILLARD, A. SIEGEL. *Integrating quantitative knowledge into a qualitative gene regulatory network*, in "PLoS Computational Biology", September 2011, vol. 7, n^{o} 9 [*DOI :* 10.1371/JOURNAL.PCBI.1002157], http://hal.archives-ouvertes.fr/hal-00626708

[3] A. BRETAUDEAU, F. COSTE, F. HUMILY, L. GARCZAREK, G. LE CORGUILLE, C. SIX, M. RATIN, O. COLLIN, W. M. SCHLUCHTER, F. PARTENSKY. *CyanoLyase: a database of phycobilin lyase sequences, motifs and functions*, in "Nucleic Acids Research", November 2012, vol. 41 [*DOI :* 10.1093/NAR/GKS1091], http://hal.inria.fr/hal-00760946

[4] F. COSTE, G. KERBELLEC. *A Similar Fragments Merging Approach to Learn Automata on Proteins*, in "ECML:Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings", J. GAMA, R. CAMACHO, P. BRAZDIL, A. JORGE, L. TORGO (editors), Lecture Notes in Computer Science, Springer, 2005, vol. 3720, pp. 522-529

[5] M. GEBSER, C. GUZIOLOWSKI, M. IVANCHEV, T. SCHAUB, A. SIEGEL, P. VEBER, S. THIELE. *Repair and Prediction (under Inconsistency) in Large Biological Networks with Answer Set Programming*, in "Principles of Knowledge Representation and Reasoning", AAAI Press, 2010

[6] C. GUZIOLOWSKI, A. BOURDÉ, F. MOREEWS, A. SIEGEL. *BioQuali Cytoscape plugin: analysing the global consistency of regulatory networks*, in "Bmc Genomics", 2009, vol. 26, n^{o} 10, 244 p. [*DOI :* 10.1186/1471-2164-10-244], http://hal.inria.fr/inria-00429804

[7] C. GUZIOLOWSKI, S. VIDELA, F. EDUATI, S. THIELE, T. COKELAER, A. SIEGEL, J. SAEZ-RODRIGUEZ. *Exhaustively characterizing feasible logic models of a signaling network using Answer Set Programming*, in "Bioinformatics", August 2013, vol. 29, n^{o} 18, pp. 2320-2326 [*DOI :* 10.1093/BIOINFORMATICS/BTT393], http://hal.inria.fr/hal-00853704

[8] J. NICOLAS, P. DURAND, G. RANCHY, S. TEMPEL, A.-S. VALIN. *Suffix-Tree Analyser (STAN): looking for nucleotidic and peptidic patterns in genomes*, in "Bioinformatics (Oxford, England)", 2005, vol. 21, pp. 4408-4410, http://hal.archives-ouvertes.fr/hal-00015234

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

[10] C. ROUSSEAU, M. GONNET, M. LE ROMANCER, J. NICOLAS. *CRISPI: a CRISPR interactive database*, in "Bioinformatics", 2009, vol. 25, n^{o} 24, pp. 3317-3318

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] A. ANTOINE-LORQUIN, A. SIEGEL, C. BELLEANNÉE. *Interest of grammatical modelling to achieve pattern matching in genomic sequences*, Université de Rennes 1, December 2016, https://hal.inria.fr/tel-01416734

[12] O. DAMERON. *Ontology-based methods for analyzing life science data*, Univ. Rennes 1, January 2016, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01403371

Articles in International Peer-Reviewed Journals

[13] V. ACUÑA, A. ARAVENA, C. GUZIOLOWSKI, D. EVEILLARD, A. SIEGEL, A. MAASS. *Deciphering transcriptional regulations coordinating the response to environmental changes*, in "BMC Bioinformatics", January 2016, vol. 17, n° 1, pp. 129-42, https://hal.archives-ouvertes.fr/hal-01260866

[14] V. BERTHÉ, J. BOURDON, T. JOLIVET, A. SIEGEL. *A combinatorial approach to products of Pisot substitutions*, in "Ergodic Theory and Dynamical Systems", 2016, vol. 36, n° 6, pp. 1757-1794 [*DOI :* 10.1017/ETDS.2014.141], https://hal.inria.fr/hal-01196326

[15] P. BORDRON, M. LATORRE, M.-P. CORTÉS, M. GONZÁLEZ, S. THIELE, A. SIEGEL, A. MAASS, D. EVEILLARD. *Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach*, in "MicrobiologyOpen", January 2016, vol. 5, n° 1, pp. 106-17, https://hal.archives-ouvertes.fr/hal-01289230

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

[17] F. J. COUSIN, S. JOUAN-LANHOUET, N. THÉRET, C. BRENNER, E. JOUAN, G. LE MOIGNE-MULLER, M.-T. DIMANCHE-BOITREL, G. JAN. *The probiotic Propionibacterium freudenreichii as a new adjuvant for TRAIL-based therapy in colorectal cancer*, in "Oncotarget", January 2016 [*DOI :* 10.18632/ONCOTARGET.6881], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01259215

[18] P. FINET, B. GIBAUD, O. DAMERON, R. LEBOUQUIN JEANNES. *Relevance of health level 7 clinical document architecture and integrating the healthcare enterprise cross-enterprise document sharing profile for managing chronic wounds in a telemedicine context*, in " Healthcare Technology Letter ", 2016, vol. 3, n° 1, pp. 22-26 [*DOI :* 10.1049/HTL.2015.0053], https://hal.archives-ouvertes.fr/hal-01301007

[19] C. GALIEZ, M. CHRISTOPHE, F. COSTE, P. BALDI. *VIRALpro: a tool to identify viral capsid and tail sequences*, in "Bioinformatics", 2016 [*DOI :* 10.1093/BIOINFORMATICS/BTV727], https://hal.archives-ouvertes.fr/hal-01242251

[20] F. GONDRET, A. VINCENT, M. HOUÉE-BIGOT, A. SIEGEL, S. LAGARRIGUE, I. LOUVEAU, D. CAUSEUR. *Molecular alterations induced by a high-fat high-fiber diet in porcine adipose tissues: variations according to the anatomical fat location*, in "BMC Genomics", December 2016, vol. 17, n° 1, article number 120 [*DOI :* 10.1186/S12864-016-2438-3], https://hal.archives-ouvertes.fr/hal-01286555

[21] M.-A. LAURENT, D. BONNIER, N. THÉRET, P. TUFFÉRY, G. MOROY. *In silico characterization of the interaction between LSKL peptide, a LAP-TGF-beta derived peptide, and ADAMTS1*, in "Computational Biology and Chemistry", April 2016, pp. 155-161 [*DOI :* 10.1016/J.COMPBIOLCHEM.2016.01.012], https://hal.inria.fr/hal-01401685

[22] G. NICOLAS, D. WALLON, C. CHARBONNIER, O. QUENEZ, S. ROUSSEAU, A.-C. RICHARD, A. ROVELET-LECRUX, S. COUTANT, K. LE GUENNEC, D. BACQ, J.-G. GARNIER, R. OLASO, A. BOLAND, V. MEYER, J.-F. DELEUZE, H. M. MUNTER, G. BOURQUE, D. AULD, A. MONTPETIT, M. LATHROP, L. GUYANT-MARÉCHAL, O. MARTINAUD, J. PARIENTE, A. ROLLIN-SILLAIRE, F. PASQUIER, I. LE BER,

M. SARAZIN, B. CROISILE, C. BOUTOLEAU-BRETONNIÈRE, C. THOMAS-ANTÉRION, C. PAQUET, M. SAUVEE, O. MOREAUD, A. GABELLE, F. SELLAL, M. CECCALDI, L. CHAMARD, F. BLANC, T. FREBOURG, D. CAMPION, D. HANNEQUIN. *Screening of dementia genes by whole-exome sequencing in early-onset Alzheimer disease: input and lessons*, in "European Journal of Human Genetics", May 2016, vol. 24, n° 5, pp. 710-716 [*DOI :* 10.1038/EJHG.2015.173], https://hal.archives-ouvertes.fr/hal-01431285

[23] M. OSTROWSKI, L. PAULEVÉ, T. SCHAUB, A. SIEGEL, C. GUZIOLOWSKI. *Boolean Network Identification from Perturbation Time Series Data combining Dynamics Abstraction and Logic Programming*, in "BioSystems", 2016 [*DOI :* 10.1016/J.BIOSYSTEMS.2016.07.009], https://hal.archives-ouvertes.fr/hal-01354075

[24] V. PICARD, O. MULNER-LORILLON, J. BOURDON, J. MORALES, P. CORMIER, A. SIEGEL, R. BELLÉ. *Model of the delayed translation of cyclin B maternal mRNA after sea urchin fertilization*, in "Molecular Reproduction and Development", 2016 [*DOI :* 10.1002/MRD.22746], http://hal.upmc.fr/hal-01390047

[25] A. ROSIER, P. MABO, L. TEMAL, P. VAN HILLE, O. DAMERON, L. DELÉGER, C. GROUIN, P. ZWEIGENBAUM, J. JACQUES, E. CHAZARD, L. LAPORTE, C. HENRY, A. BURGUN. *Personalized and automated remote monitoring of atrial fibrillation*, in "EP-Europace", 2016, vol. 18, n° 3, pp. 347-352 [*DOI :* 10.1093/EUROPACE/EUV234], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01331019

[26] A. ROSIER, P. MABO, L. TEMAL, P. VAN HILLE, O. DAMERON, L. DELÉGER, C. GROUIN, P. ZWEIGENBAUM, J. JACQUES, E. CHAZARD, L. LAPORTE, C. HENRY, A. BURGUN. *Remote Monitoring of Cardiac Implantable Devices: Ontology Driven Classification of the Alerts*, in "Studies in Health Technology and Informatics", 2016, vol. 221, pp. 59–63 [*DOI :* 10.3233/978-1-61499-633-0-59], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01319949

[27] S. SWARUP SAMAL, A. NALDI, D. GRIGORIEV, A. WEBER, N. THÉRET, O. RADULESCU. *Geometric analysis of pathways dynamics: application to versatility of TGF-β receptors*, in "BioSystems", July 2016, vol. 149, pp. 3–14 [*DOI :* 10.1016/J.BIOSYSTEMS.2016.07.004], https://hal.inria.fr/hal-01379033

[28] S. VIDELA, J. SAEZ-RODRIGUEZ, C. GUZIOLOWSKI, A. SIEGEL. *caspo: a toolbox for automated reasoning on the response of logical signaling networks families*, in "Bioinformatics", 2016 [*DOI :* 10.1093/BIOINFORMATICS/BTW738], https://hal.inria.fr/hal-01426880

International Conferences with Proceedings

[29] T. GUYET, Y. MOINARD, J. NICOLAS, R. QUINIOU. *Packing graphs with ASP for landscape simulation*, in "IJCAI 2016 - 25th International joint conference on artificial intelligence", New-york, United States, AAAI, July 2016, 8 p. , https://hal.inria.fr/hal-01327368

National Conferences with Proceedings

[30] Y. RIVAULT, N. LE MEUR, O. DAMERON. *Complications post-opératoires et mode de prise en charge en angioplastie : apport du Programme de Médicalisation des Systèmes d'Information (PMSI)*, in "Congrès Adelf-Epiter 2016", Rennes, France, September 2016, https://hal.inria.fr/hal-01389688

[31] Y. RIVAULT, N. LE MEUR, O. DAMERON. *La gestion de données médico-administratives grâce aux outils du Web sémantique*, in "Journées ADELF-EMOIS « Système d'information hospitalier et Epidémiologie »", Dijon, France, Revue d'Épidémiologie et de Santé Publique, March 2016, vol. 64, n° Supplement 1, S15 p. [*DOI :* 10.1016/J.RESPE.2016.01.051], https://hal.archives-ouvertes.fr/hal-01297389

Conferences without Proceedings

[32] A. ANTOINE-LORQUIN, F. MAHÉ, M. DUNTHORN, C. BELLEANNÉE. *Impact de la recherche d'amorces mutées sur les résultats d'analyses métagénomiques*, in "Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)", Rennes, France, Jobim-2016, Société Française de Bioinformatique (SFBI), June 2016, https://hal.archives-ouvertes.fr/hal-01343121

[33] F. BALUSSON, M.-A. BOTREL, O. DAMERON, Y. DAUXAIS, E. DREZEN, A. DUPUY, T. GUYET, D. GROSS-AMBLARD, A. HAPPE, N. LE MEUR, B. LE NAUTOUT, E. LERAY, E. NOWAK, C. RAULT, E. OGER, E. POLARD. *PEPS: a platform for supporting studies in pharmaco-epidemiology using medico-administrative databases*, in "International Congress on e-Health Research", Paris, France, October 2016, https://hal.inria.fr/hal-01380939

[34] M. CHEVALLIER, M. AITE, J. GOT, G. COLLET, N. LOIRA, M.-P. CORTÉS, C. FRIOUX, J. LANIAU, C. TROTTIER, A. MAASS, A. SIEGEL. *Handling the heterogeneity of genomic and metabolic networks data within flexible workflows with the PADMet toolbox*, in "Jobim 2016 : 17ème Journées Ouvertes en Biologie, Informatique et Mathématiques", Lyon, France, June 2016, https://hal.inria.fr/hal-01377844

[35] F. COSTE, M. DEMIRDELEN. *A refined parsing graph approach to learn smaller contextually substitutable grammars with less data*, in "ICGI 2016 - 13th International Conference on Grammatical Inference", Delft, Netherlands, October 2016, https://hal.inria.fr/hal-01406337

[36] W. DYRKA, F. COSTE, O. UNOLD, L. CULER, A. KACZMAREK. *How to measure the topological quality of protein grammars?*, in "ICGI 2016 - 13th International Conference on Grammatical Inference", Delft, Netherlands, October 2016, https://hal.inria.fr/hal-01406331

[37] F. LEGEAI, C. BETTEMBOURG, A. BRETAUDEAU, Y. CHAUSSIN, O. DAMERON, D. TAGU. *BIPAA/Askomics, a new and easy approach for querying genomics and epigenomics elements in interaction*, in "XXVth International Congress of Entomology 2016,", Orlando, Florida, United States, September 2016, https://hal.inria.fr/hal-01391080

[38] F. MOREEWS, O. SALLOU, O. COLLIN. *An application suite based on the IFB Container as a Service platform*, in "ECCB 2016", Den Haag, Netherlands, ISCB, September 2016, https://hal.inria.fr/hal-01394295

[39] A. SIEGEL. *A system biology loop for the identification of boolean networks to model signaling networks.* , in "European Conference on Mathematical and Theoretical Biology", Nottingham, France, 2016, https://hal.inria.fr/hal-01399456

Scientific Books (or Scientific Book chapters)

[40] F. COSTE. *Learning the Language of Biological Sequences*, in "Topics in Grammatical Inference", J. HEINZ, J. M. SEMPERE (editors), Springer-Verlag, 2016 [*DOI :* 10.1007/978-3-662-48395-4_8], https://hal.inria.fr/hal-01244770

Other Publications

[41] F. COUSIN, S. JOUAN-LANHOUET, N. THÉRET, C. BRENNER, E. JOUAN, G. LE MOIGNE-MULLER, M.-T. DIMANCHE-BOITREL. *The probiotic [i]Propionibacteriumfreudenreichii[/i] as a new adjuvant for TRAIL-based therapy in colorectal cancer*, September 2016, 4 . International Symposium on Propionibacteria and Bibfidobacteria , Poster, https://hal.archives-ouvertes.fr/hal-01372131

[42] F. COUSIN, S. JOUAN-LANHOUET, N. THÉRET, C. BRENNER, E. JOUAN, G. LE MOIGNE-MULLER, M.-T. DIMANCHE-BOITREL, G. JAN. *The probiotic [i]Propionibacterium freudenreichii[/i] as a new adjuvant for TRAIL-based therapy in colorectal cancer*, June 2016, International Scientific Conference on Probiotics and Prebiotics - IPC2016, Poster, https://hal.archives-ouvertes.fr/hal-01286497

[43] M. DEMIRDELEN. *Fast parser for biological sequences and a new algorithm for the inference of substitutable languages*, Univ. Rennes 1, June 2016, https://hal.inria.fr/hal-01406352

[44] A. EVRARD, C. BETTEMBOURG, O. DAMERON, O. FILANGI, A. BRETAUDEAU, F. F. LEGEAI, R. DELOURME, M. MANZANARES-DAULEUX, M. JUBAULT. *Comparative genomic analysis of Clubroot resistance in the Brassicaceae family*, October 2016, Brassica2016, Poster, https://hal.inria.fr/hal-01391707

[45] A. EVRARD, C. BETTEMBOURG, M. M. JUBAULT, O. DAMERON, O. FILANGI, A. BRETAUDEAU, F. F. LEGEAI. *Integration and query of biological datasets with Semantic Web technologies: AskOmics*, June 2016, Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2016), Poster, https://hal.inria.fr/hal-01391087

[46] S. RÉMY. *Apprentissage profond et acquisition de représentations latentes de séquences peptidiques*, ENS Rennes, August 2016, https://hal.inria.fr/hal-01406368

References in notes

[47] O. ABDOU-ARBI, S. LEMOSQUET, J. VAN MILGEN, A. SIEGEL, J. BOURDON. *Exploring metabolism flexibility in complex organisms through quantitative study of precursor sets for system outputs*, in "BMC Systems Biology", 2014, vol. 8, n^{o} 1, 8 p. [*DOI :* 10.1186/1752-0509-8-8], https://hal.inria.fr/hal-00947219

[48] V. ACUÑA, A. ARAVENA, A. MAASS, A. SIEGEL. *Modeling parsimonious putative regulatory networks: complexity and heuristic approach*, in "15th conference in Verification, Model Checking, and Abstract Interpretation", San Diego, United States, Springer, 2014, vol. 8318, pp. 322-336 [*DOI :* 10.1007/978-3-642-54013-4_18], https://hal.inria.fr/hal-00926477

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

[50] C. BARAL. *Knowledge Representation, Reasoning and Declarative Problem Solving*, Cambridge University Press, 2010

[51] T. BAUMURATOVA, D. SURDEZ, B. DELYON, G. STOLL, O. DELATTRE, O. RADULESCU, A. SIEGEL. *Localizing potentially active post-transcriptional regulations in the Ewing's sarcoma gene regulatory network*, in "BMC Systems Biology", 2010, vol. 4, n^{o} 1, 146 p. [*DOI :* 10.1186/1752-0509-4-146], http://www.hal.inserm.fr/inserm-00984711

[52] R. BELLAZZI. *Big Data and Biomedical Informatics: A Challenging Opportunity*, in "Yearbook of medical informatics", 2014, vol. 9, n^{o} 1, In press

[53] R. BELLAZZI, M. DIOMIDOUS, I. N. SARKAR, K. TAKABAYASHI, A. ZIEGLER, A. T. MCCRAY. *Data analysis and data mining: current issues in biomedical informatics*, in "Methods of information in medicine", 2011, vol. 50, n° 6, pp. 536–544

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

[55] R. BELLÉ, S. PRIGENT, A. SIEGEL, P. CORMIER. *Model of cap-dependent translation initiation in sea urchin: a step towards the eukaryotic translation regulation network*, in "Molecular Reproduction and Development", 2010, vol. 77, n° 3, pp. 257-64

[56] R. BENDAOUD, A. NAPOLI, Y. TOUSSAINT. *Formal Concept Analysis: A Unified Framework for Building and Refining Ontologies*, in "Knowledge Engineering: Practice and Patterns, 16th International Conference, EKAW 2008, Acitrezza, Italy, September 29 - October 2, 2008. Proceedings", A. GANGEMI, J. EUZENAT (editors), Lecture Notes in Computer Science, Springer, 2008, vol. 5268, pp. 156-171, http://dx.doi.org/10.1007/978-3-540-87696-0_16

[57] C. BETTEMBOURG, C. DIOT, O. DAMERON. *Optimal threshold determination for interpreting semantic similarity and particularity*, in "PLoS ONE", 2015, vol. 10, n° 7, e0133579 [*DOI :* 10.1371/JOURNAL.PONE.0133579], https://hal.archives-ouvertes.fr/hal-01207763

[58] C. BETTEMBOURG, C. DIOT, O. DAMERON. *Optimal Threshold Determination for Interpreting Semantic Similarity and Particularity: Application to the Comparison of Gene Sets and Metabolic Pathways Using GO and ChEBI*, in "PLoS ONE", 2015, 30 p. [*DOI :* 10.1371/JOURNAL.PONE.0133579], https://hal.inria.fr/hal-01184934

[59] C. BIZER, T. HEATH, T. BERNERS LEE. *Linked Data–The story so far*, in "International Journal on Semantic Web and Information Systems", 2009, vol. 5, n° 3, pp. 1–22

[60] J. A. BLAKE, C. J. BULT. *Beyond the data deluge: Data integration and bio-ontologies*, in "Journal of Biomedical Informatics", 2006, vol. 39, n° 3, pp. 314–320

[61] P. BLAVY, F. GONDRET, S. LAGARRIGUE, J. VAN MILGEN, A. SIEGEL. *Using a large-scale knowledge database on reactions and regulations to propose key upstream regulators of various sets of molecules participating in cell metabolism*, in "BMC Systems Biology", 2014, vol. 8, n° 1, 32 p. [*DOI :* 10.1186/1752-0509-8-32], https://hal.inria.fr/hal-00980499

[62] P. BORDRON, D. EVEILLARD, A. MAASS, A. SIEGEL. *An ASP application in integrative biology: identification of functional gene units*, in "LPNMR - 12th Conference on Logic Programming and Nonmonotonic Reasoning - 2013", Corunna, Spain, September 2013, http://hal.inria.fr/hal-00853762

[63] H. CHEN, T. YU, J. Y. CHEN. *Semantic Web meets Integrative Biology: a survey*, in "Briefings in bioinformatics", 2012, vol. 14, n° 1, pp. 109–125

[64] G. COLLET, D. EVEILLARD, M. GEBSER, S. PRIGENT, T. SCHAUB, A. SIEGEL, S. THIELE. *Extending the Metabolic Network of Ectocarpus Siliculosus using Answer Set Programming*, in "LPNMR - 12th Conference

on Logic Programming and Nonmonotonic Reasoning - 2013", Corunna, Spain, September 2013, http://hal.inria.fr/hal-00853752

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

[66] S. DAMINELLI, V. J. HAUPT, M. REIMANN, M. SCHROEDER. *Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network*, in "Integr. Biol.", 2012, vol. 4, pp. 778-788, http://dx.doi.org/10.1039/C2IB00154C

[67] O. DEMEURE, F. LECERF, C. DUBY, C. DESERT, S. DUCHEIX, H. GUILLOU, S. LAGARRIGUE. *Regulation of LPCAT3 by LXR*, in "Gene", Jan 2011, vol. 470, n^{o} 1-2, pp. 7–11

[68] S. M. DITTAMI, T. BARBEYRON, C. BOYEN, J. CAMBEFORT, G. COLLET, L. DELAGE, A. GOBET, A. GROISILLIER, C. LEBLANC, G. MICHEL, D. SCORNET, A. SIEGEL, J. E. TAPIA, T. TONON. *Genome and metabolic network of "Candidatus Phaeomarinobacter ectocarpi" Ec32, a new candidate genus of Alphaproteobacteria frequently associated with brown algae*, in "Frontiers in Genetics", 2014, vol. 5, 241 p. [*DOI :* 10.3389/FGENE.2014.00241], https://hal.inria.fr/hal-01079739

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

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

[71] F. GONDRET, A. VINCENT, M. HOUEE, S. LAGARRIGUE, A. SIEGEL, D. CAUSEUR, I. LOUVEAU. *Integrative responses of pig adipose tissues to high-fat high-fiber diet: towards key regulators of energy flexibility*, March 2015, ASAS/ADSA midwest meeting, Poster, https://hal.archives-ouvertes.fr/hal-01210925

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

[73] R. KAMINSKI, A. SIEGEL, T. SCHAUB, S. VIDELA. *Minimal Intervention Strategies in Logical Signaling Networks with ASP*, in "Theory and Practice of Logic Programming", September 2013, vol. 13, n^{o} Special issue 4-5, pp. 675-690 [*DOI :* 10.1017/S1471068413000422], http://hal.inria.fr/hal-00853747

[74] S. LAURENT, A. RICHARD, O. MULNER-LORILLON, J. MORALES, D. FLAMENT, V. GLIPPA, J. BOURDON, P. GOSSELIN, A. SIEGEL, P. CORMIER, R. BELLÉ. *Modelization of the regulation of protein synthesis following fertilization in sea urchin shows requirement of two processes: a destabilization of eIF4E:4E-BP complex and a great stimulation of the 4E-BP-degradation mechanism, both rapamycin-sensitive*, in "Frontiers in Genetics", 2014, vol. 5, 117 p. [*DOI :* 10.3389/FGENE.2014.00117], https://hal.inria.fr/hal-01079758

[75] F. LEGEAI, T. DERRIEN, V. WUCHER, D. AUDREY, G. LE TRIONNAIRE, D. TAGU. *Long non-coding RNA in the pea aphid; identification and comparative expression in sexual and asexual embryos*, in "Arthropod Genomics Symposium", Urbana, United States, June 2014, https://hal.inria.fr/hal-01091304

[76] S.-W. LEUNG, C. MELLISH, D. ROBERTSON. *Basic Gene Grammars and DNA-ChartParser for language processing of Escherichia coli promoter DNA sequences*, in "Bioinformatics", 2001, vol. 17, n^o 3, pp. 226-236 [*DOI :* 10.1093/BIOINFORMATICS/17.3.226], http://bioinformatics.oxfordjournals.org/content/17/3/226.abstract

[77] A. LIHONOSOVA, A. KAMINSKAYA. *Using Formal Concept Analysis for Finding the Closest Relatives among a Group of Organisms*, in "Procedia Computer Science", 2014, vol. 31, n^o Complete, pp. 860-868, http://dx.doi.org/10.1016/j.procs.2014.05.337

[78] K. M. LIVINGSTON, M. BADA, W. A. BAUMGARTNER, L. E. HUNTER. *KaBOB: ontology-based semantic integration of biomedical databases*, in "BMC bioinformatics", 2015, vol. 16, 126 p.

[79] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAIL-TABBONE. *Many-Valued Concept Lattices for Conceptual Clustering and Information Retrieval*, in "Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence", Amsterdam, The Netherlands, The Netherlands, IOS Press, 2008, pp. 127–131, http://dl.acm.org/citation.cfm?id=1567281.1567313

[80] M. OBITKO, V. SNÁSEL, J. SMID. *Ontology Design with Formal Concept Analysis*, in "Proceedings of the CLA 2004 International Workshop on Concept Lattices and their Applications, Ostrava, Czech Republic, September 23-24, 2004", V. SNÁSEL, R. BELOHLÁVEK (editors), CEUR Workshop Proceedings, CEUR-WS.org, 2004, vol. 110, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-110/paper12.pdf

[81] P.-F. PLUCHON, T. FOUQUEAU, C. CREZE, S. LAURENT, J. BRIFFOTAUX, G. HOGREL, A. PALUD, G. HENNEKE, A. GODFROY, W. HAUSNER, M. THOMM, J. NICOLAS, D. FLAMENT. *An Extended Network of Genomic Maintenance in the Archaeon Pyrococcus abyssi Highlights Unexpected Associations between Eucaryotic Homologs*, in "PLoS ONE", 2013, vol. 8, n^o 11, e79707 [*DOI :* 10.1371/JOURNAL.PONE.0079707], http://hal.inria.fr/hal-00911795

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

[83] M. ROUANE-HACENE, M. HUCHARD, A. NAPOLI, P. VALTCHEV. *Using Formal Concept Analysis for Discovering Knowledge Patterns*, in "CLA'10: 7th International Conference on Concept Lattices and Their Applications", Sevilla, Spain, S. O. MARZENA KRYSZKIEWICZ (editor), CEUR, University of Sevilla, October 2010, n^o 672, pp. 223-234, http://hal-lirmm.ccsd.cnrs.fr/lirmm-00531802

[84] C. ROUSSEAU, M. GONNET, M. LE ROMANCER, J. NICOLAS. *CRISPI: a CRISPR interactive database*, in "Bioinformatics", 2009, vol. 25, n^o 24, pp. 3317-3318

[85] L. ROYER, M. REIMANN, B. ANDREOPOULOS, M. SCHROEDER. *Unraveling Protein Networks with Power Graph Analysis*, in "PLoS Comput Biol", 07 2008, vol. 4, n^o 7, e1000108, http://dx.plos.org/10.1371%2Fjournal.pcbi.1000108

[86] T. SCHAUB, A. SIEGEL, S. VIDELA. *Reasoning on the response of logical signaling networks with Answer Set Programming*, in "Logical Modeling of Biological Systems", Wiley Online Librairy, 2014, pp. 49-92 [*DOI :* 10.1002/9781119005223.CH2], https://hal.inria.fr/hal-01079762

[87] T. SCHAUB, S. THIELE. *Metabolic Network Expansion with Answer Set Programming*, in "ICLP 2009", LNCS, Springer, 2009, vol. 5649, pp. 312-326

[88] S. SCHULZ, L. BALKANYI, R. CORNET, O. BODENREIDER. *From Concept Representations to Ontologies: A Paradigm Shift in Health Informatics?*, in "Healthcare informatics research", 2013, vol. 19, n^{o} 4, pp. 235-242

[89] D. SEARLS. *The language of genes*, in "Nature", 2002, vol. 420, pp. 211-217

[90] D. SEARLS. *String Variable Grammar: A Logic Grammar Formalism for the Biological Language of DNA*, in "Journal of Logic Programming", 1995, vol. 24, n^{o} 1&2, pp. 73-102

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

[92] S. TEMPEL, C. ROUSSEAU, F. TAHI, J. NICOLAS. *ModuleOrganizer: detecting modules in families of transposable elements*, in "BMC Bioinformatics", 2010, vol. 11, 474 p. [*DOI :* 10.1186/1471-2105-11-474], http://hal.inria.fr/inria-00536742

[93] S. THIELE, J. SAEZ-RODRIGUEZ, L. CERONE, A. SIEGEL, C. GUZIOLOWSKI, S. KLAMT. *Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies*, in "BMC Bioinformatics", 2015, vol. 16, 345 p. [*DOI :* 10.1186/S12859-015-0733-7], https://hal.inria.fr/hal-01225228

[94] S. VIDELA, C. GUZIOLOWSKI, F. EDUATI, S. THIELE, M. GEBSER, J. NICOLAS, J. SAEZ-RODRIGUEZ, T. SCHAUB, A. SIEGEL. *Learning Boolean logic models of signaling networks with ASP*, in "Journal of Theoretical Computer Science (TCS)", September 2015, n^{o} 599, pp. 79-101 [*DOI :* 10.1016/J.TCS.2014.06.022], https://hal.inria.fr/hal-01058610

[95] S. VIDELA, I. KONOKOTINA, L. ALEXOPOULOS, J. SAEZ-RODRIGUEZ, T. SCHAUB, A. SIEGEL, C. GUZIOLOWSKI. *Designing experiments to discriminate families of logic models*, in "Frontiers in Bioengineering and Biotechnology", September 2015, 9 p. [*DOI :* 10.3389/FBIOE.2015.00131], https://hal.inria.fr/hal-01196178

[96] R. WILLE. *restructuring lattice theory: an approach based on hierarchies of concepts*, in "Proceedings of the 7th International Conference on Formal Concept Analysis", Berlin, Heidelberg, ICFCA '09, Springer-Verlag, 2009, pp. 314–339, http://dx.doi.org/10.1007/978-3-642-01815-2_23

[97] V. WUCHER, D. TAGU, J. NICOLAS. , B. LAUSEN, S. KROLAK-SCHWERDT, M. BÖHMER (editors) *Edge Selection in a Noisy Graph by Concept Analysis – Application to a Genomic Network*, Data Science, Learning by Latent Structures, and Knowledge Discovery, Springer, 2014, 550 p. , https://hal.inria.fr/hal-01093337

[98] V. WUCHER. *Modeling of a gene network between mRNAs and miRNAs to predict gene functions involved in phenotypic plasticity in the pea aphid*, Universite Rennes 1, November 2014, https://hal.archives-ouvertes.fr/tel-01095967