IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Activity Report 2013

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

Keywords: Computational Biology, Systems Biology, Knowledge Representation, Machine Learning, Combinatorial Optimization, Formal Methods, Network Modeling

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

1. Members

Research Scientists
Anne Siegel [Team leader, CNRS, Senior Researcher, HdR]
François Coste [Inria, Researcher]
Jacques Nicolas [Inria, Senior Researcher, HdR]

Faculty Members
Catherine Belleannée [Univ. Rennes I, Associate Professor]
Olivier Dameron [Univ. Rennes I, Associate Professor]
Laurent Miclet [Univ. Rennes I, Emerite Professor, HdR]
Michel Le Borgne [Univ. Rennes I, until Sep 2013]

Engineers
Jeanne Cambefort [CNRS, from Mar 2013]
Guillaume Collet [CNRS]
Claudia Hériveau [Inria, until Nov 2013]
Olivier Quenez [CNRS]

PhD Students
Oumarou Abdou-Arbi [Univ. Rennes I, until Oct 2013]
Geoffroy Andrieux [Univ. Rennes I, until Aug 2013]
Aymeric Antoine-Lorquin [Univ. Rennes I, from Jan. 2013]
Andres Aravena [Univ. Chile, until Dec 2013]
Charles Bettembourg [Univ. Rennes I, from Apr 2013 until Dec 2013]
Clovis Galiez [Inria]
Gaëlle Garet [Inria]
Julie Laniau [Inria, from Oct 2013]
Vincent Picard [Univ. Rennes I]
Sylvain Prigent [Univ. Rennes I]
Santiago Videla [CNRS]
Valentin Wucher [INRA]

Post-Doctoral Fellow
Sven Thiele [Inria, until Sep 2013]

Administrative Assistant
Marie-Noëlle Georgeault [Inria]

Others
Jérémie Bourdon [Associated Collaborator, Univ. Nantes]
Damien Eveillard [Associated Collaborator, Univ. Nantes]

2. Overall Objectives

2.1. Overall objectives

The research domain of the Dyliss team is bioinformatics and systems biology. Our main goal in biology is to characterize groups of genetic actors that control the phenotypic answer of non-model species when challenged by their environment. Unlike model species, a limited prior-knowledge is available for these organisms together with a small range of experimental studies (culture conditions, genetic transformations). To overcome these limitations, 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 take into account both the information on physiological responses of the studied species under various constraints and the genetic information from their long-distant cousins.

The challenge to face is thus incompleteness: limited range of physiological or genetic known perturbations together with an incomplete knowledge of living mechanisms involved. 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, the biologist will be in position of developing experimental strategies to progressively shrink the space of hypotheses and gain in the understanding of the system. This refinement approach is particularly suited to non-model organisms, which have specific and little known survival mechanisms. It is also required in the framework of an increasing automation of experimentations in biology.

From the bioinformatics aspect, the main challenge is to transfer genome-level information available in well-annotated organisms on their distant relatives. To that matter, we develop methods within the context of formal systems to identify and formalize the genomic specificities of target species which are observed at the physiological level rather than at the genome-level. Our main purpose is to combine in a suitable way machine learning, logical constraints and dynamical systems techniques to get a combinatorial representation of the space of admissible models for groups of genome products implied in the answer of the species. The steps of the analysis are to (*i*) formalize and integrate in a set of logic constraints the genetic information and the physiological responses; (*ii*) investigate the space of admissible models and exhibit its structure and main features; (*iii*) identify corresponding genomic products within sequences.

We target applications in marine biology and environmental microbiology, that is, organisms with a good long-term biotechnological potential but requiring prior intensive in-silico studies to fully exploit their specificities. We focus on unicellular and pluricellular organisms with a relatively simple development but very specific physiological capabilities. 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") whereas in environmental microbiology we collaborate both with the CRG in Chile in the framework of the Ciric Chilean inria center (*Ciric-Omics*) and with laboratories in Rennes (Inra).

2.2. Highlights of the Year

The collaboration with Universidad de Chile was strengthened by the organization of a workshop in Chile gathering Chilean, French and German partners about the modeling of biological systems [website], the defense of a co-supervised Ph-D thesis [13], and a graduate-level course given by a Dyliss member in Chile.

3. Research Program

3.1. Knowledge representation with constraint programming

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. In the last few years, our team has investigated an alternative perspective where each observation 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 need to relax the problem and to study the space of corrections to prior knowledge in order to fit reasonably with observation data. In this case, this issue is modeled as combinatorial (sub)-optimization issues. In both cases, 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 satisfiability problem (in the case of existing solutions) or to the optimization problem (in the case of required corrections to the prior knowledge) [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** [36], [41] 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 [1], and proposed robust corrections [5]. See Fig. 1 for details. The results obtained so far suggest that this approach is compatible with efficiency, scale and expressivity needed by biological systems. Our goal is now to provide **formal models of queries on biological networks** with the focus of integrating dynamical information as explicit logical constraints in the modeling process. This would definitely introduce such logical paradigms as a powerful approach to build and query reconstructed biological systems, in complement to discriminative approaches. 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 whether progresses in the field of constraint logical programming, shown by the performance of ASP-solvers in several recent competitions, are now sufficient to address the complexity of constraint-satisfiability and combinatorial optimization issues explored in systems biology.

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 the models that best fit biological knowledge. 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.

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 opens 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.

3.2. Probabilistic and symbolic dynamics

We work on new techniques to emphasize biological strategies that must occur to reproduce quantitative measurements in order to predict the quantitative response of a system at a larger-scale. Our framework mixes mechanistic and probabilistic modeling [2]. The system is modeled by an Event Transition Graph, that is, a **Markovian qualitative description of its dynamics** together with quantitative laws which describe the effect of the dynamic transitions over higher scale quantitative measurements. Then, a few time-series quantitative

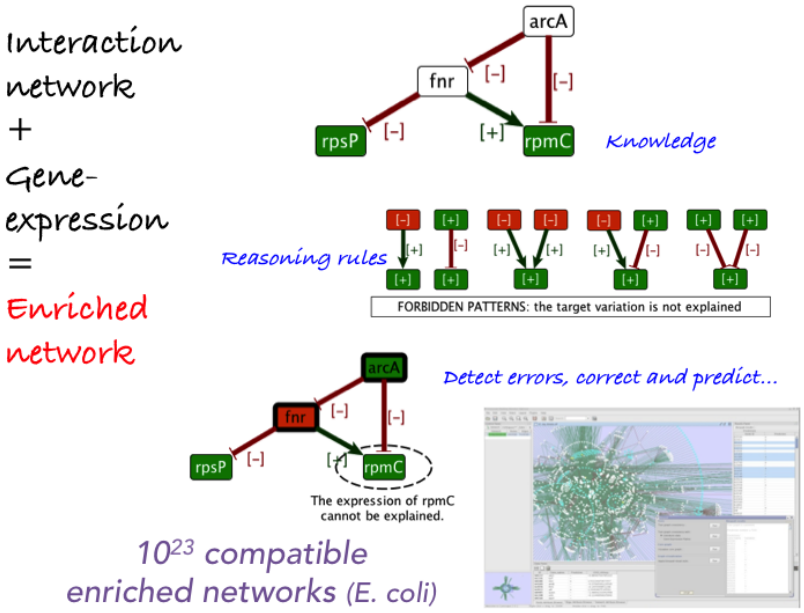

Figure 1.

An example of reasoning process in order to identify which expression of non-observed nodes (white nodes) are fixed by partial observations and rules derived from the system dynamics. The ASP-based logical approach is flexible enough to model in a single framework network characteristics (products, interactions, partial information on signs of regulations and observations) and static rules about the effects of the dynamics of the system. Extensions of this framework include the exhaustive search for system repair or more constrained dynamical rules. [6], [5] ***Step 1*** *. Regulation knowledge is represented as a signed oriented graph. Edge colors stand for regulatory effects (red/green→ inhibition or activation). Vertex colors stand for gene expression data (red/green→ under or over-expression).* ***Step 2*** *. Integrity constraints on the whole colored graph come from the necessity to find a consistent explanation of the link between regulation and expression.* ***Step 3*** *. The model allows both the prediction of values (e.g. for fnr in the figure) and the detection of contradictions (e.g. the expression level of rpmC is inconsistant with the regulation in the graph). .*

measurements are provided. Following an ergodic assumption and average case analysis properties, we know that a multiplicative accumulation law on a Markov chain asymptotically follows a log-normal law with explicit parameters [40]. This property can be derived into constraints to describe the set of admissible weighted Markov chains whose asymptotic behavior agrees with the quantitative measures at hand. A precise study of this constrained space via local search optimization emphasizes the most important discrete events that must occur to reproduce the information at hand. These methods have been validated on the *E. coli* regulatory network benchmark. See Figure 2 for illustration. We now plan to apply these techniques to reduced networks representing the main pathways and actors automatically generated from the integrative methods developed in the former section. This requires to improve the range of dynamics that can be modeled by these techniques, as well as the efficiency and scalability of the local search algorithms.

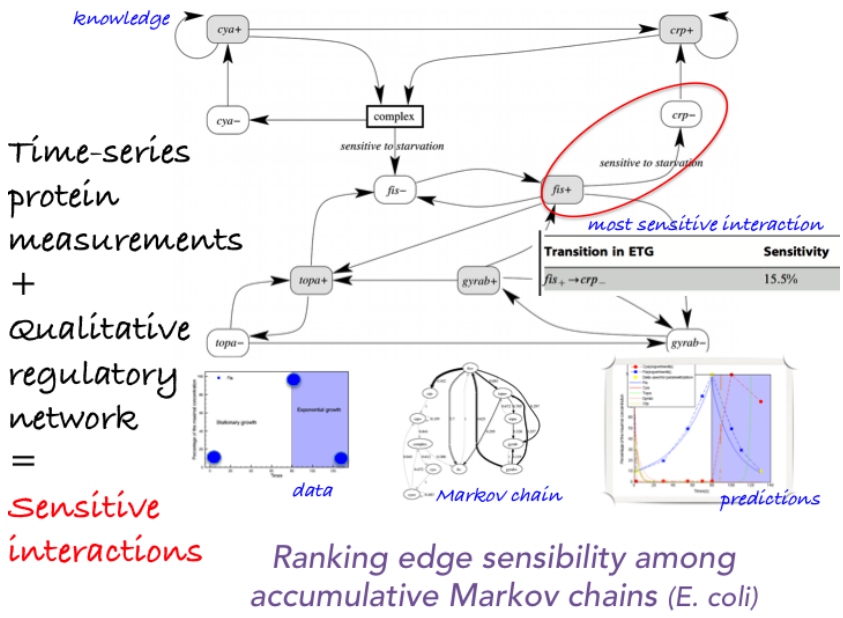

Figure 2.

Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems and identification of key interactions [2].
Input data *are provided by a qualitative description of the system dynamics at the transcription level (interaction graph) and 3 concentration measurements of the* fis *protein (population scale). The method computes an* ***Event-Transition Graph****. Interaction frequencies required to predict he population scale behavior as the asymptotic behavior of an accumulation multiplicative law over a Markov chain. Estimation by local searches in the space of Markov chains consistent with the observed dynamics and whose asymptotic behavior is consistent with quantitative observations at the population scale. Edge thickness reflects their sensitivity in the search space. It allows to* ***predict*** *the* Cya *protein concentration (red curve) which fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in E. Coli response to nutritional stress.*

3.3. Grammatical inference and highly expressive structures

Our main field of expertise in machine learning concerns grammatical models with a strong expertise in finite state automata learning. By introducing a similar fragment merging heuristic approach, 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. As an example, this tools allows us to properly model the multi-domain function of the protein family TNF, which is impossible with other existing probabilistic-based approach (see Fig. 3). It was also applied to model families of proteins in cyanobacteria [3]. Our future goal is to demonstrate the relevance of formal language theory by addressing the question of enzyme prediction, from their genomic or protein sequences, aiming at better sensitivity and specificity. 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 is 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. We wish to extend our expertise towards inferring Context-Free Grammars including the topological information coming from the structural characterization of active sites.

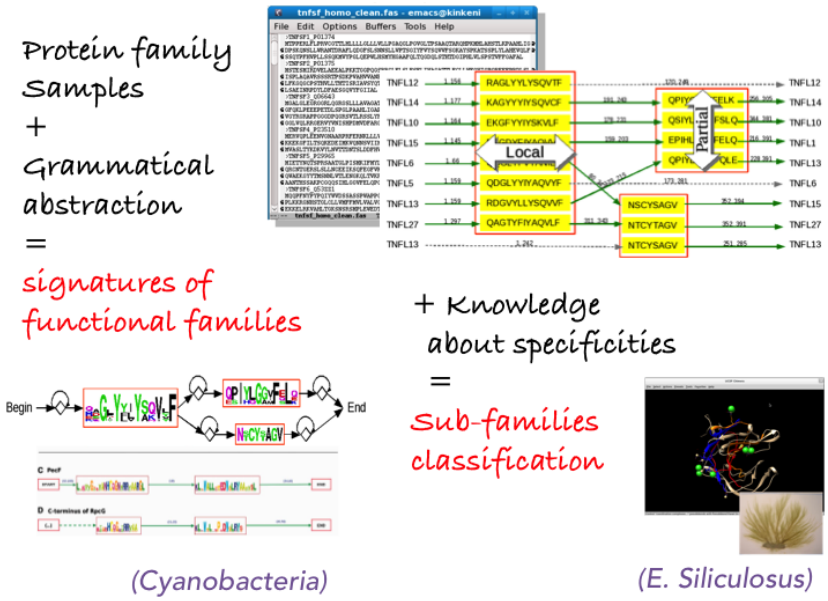

Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences, a partial local alignment is computed and an automaton is inferred, which can be considered as a signature of the family of proteins. This allows searching for new members of the family [3]. Adding further information about the specific properties of proteins within the family allows to exhibit a refined classification.*

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 [42]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [46], [45]. This increases the expressivity of the formalism towards midly 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. We have designed a tool, STAN (suffix-tree analyser) which makes it possible to search for a subset of SVG patterns in full chromosome sequences [9]. This tool

was used for the recognition of transposable elements in *Arabidopsis thaliana* [47] or for the design of a CRISPR database [10]. See Figure 4 for illustration. Our goal is to extend the framework of STAN. 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. In this direction, we are working on Logol, a language and framework based on a systematic introduction of constraints on string variables.

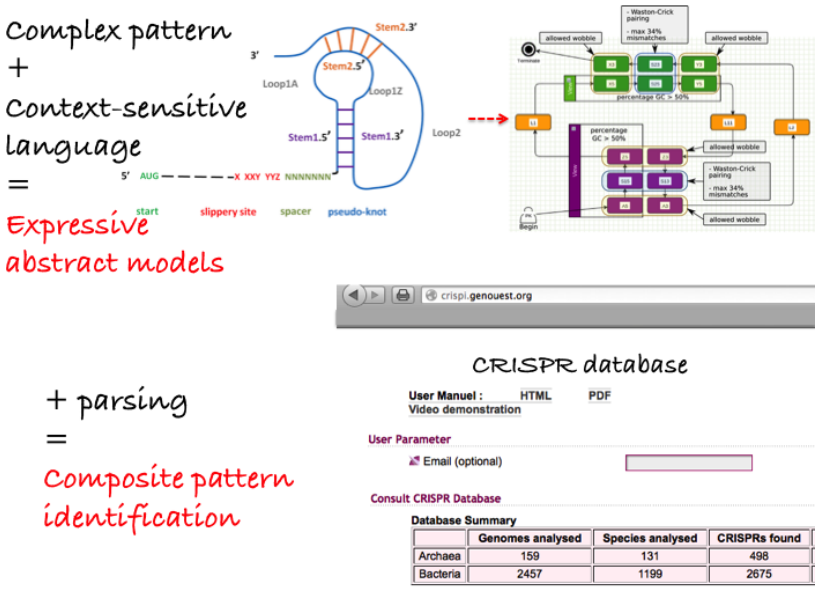

Figure 4. A typical RNA structure such as the pseudo-knot can be graphical modeling of a pseudo-knot based on the expressivity of String Variable Grammars used in the Logol framework. Combined with parsers, this leads to composite pattern identification such as CRISPR [43].

4. Application Domains

4.1. Formal models in molecular biology

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.

Biological data integration. 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 implied 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.

Asymptotic dynamics of a biological system 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 Pufa's are found to be produced using a set of pathways specific to brown algae, the work in axis 2 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 Pufa's 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 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 Pufa's controlling pathways have to be precisely identified. Actors structures are represented by syntactic models (see Fig. 4). 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.

4.2. Application fields

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 [48] [27]. On the same application model, the pattern discovery tool *protomata learner* allows for the classification of sub-families of specific proteins. The same tool also allowed us to gain a better understanding of cyanobacteria proteins [3]. Finally, in dynamical systems, we use asymptotic analysis (tool *pogg*) to decipher the initiation of sea urchin translation [37]. We are currently initiating two new research programs in this domain: the team will participate to a collaboration program with the Biocore and Ange Inria teams, focused on the understanding on green micro-algae; and we will be involved in the deciphering of phytoplancton variability at the system biology level in collaboration with the Station Biologique de Roscoff.

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 *bioquali, ingranalysis, shogun, lombarde* were designed in this context [6] [26]. In parallel, in collaboration with Ifremer (Brest), we have conducted similar work to decipher protein-protein interactions within archebacteria [20]. Our sequence analysis tool (*logol*) allowed us to build and maintain a very expressive CRISPR database [10].

Similarly, in **animal biology**, our goal is to propose methods to identify regulators of very complex phenotypes related to nutritional issues. In collaboration with researchers from Inra/Pegase and Inra/Igeep laboratories, we develop methods to distinguish the response of cows, chicken or porks to different diaries or treatments and characterize upstream transcriptional regulators for such a response. The system biology tool *nutritional analyzer* was designed in this framework [15]. The pattern matching tool *logol* also allows for a fine identification of transcription factor motifs [39]. Constraints-based programming also allows us to decipher regulators of reproduction for pea aphids.

We are less involved in **health** as the models and data studied in this application field are well informed and rather data-driven. 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 [1]. 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 [18], [19] and N. Theret group at Inserm/Irset (Rennes) [35]. The dynamical system tools *caspo and cadbiom* were designed within these collaborations. Future studies will focus on the understanding of the metabolism of xenobiotics, still in collaboration with Inserm/Irset (Rennes).

5. Software and Platforms

5.1. Platform

Our tools are based on formal systems. They aim at guiding the user to progressively reduce the space of models (gene or protein families, set of main actors involved in a system response, dynamical models) which are compatible with both knowledge and experimental observations. Most of our tools are available both as stand-alone software and through portals such as Mobyle or Galaxy interfaces. Tools are developed in collaboration with the GenOuest resource and data center hosted in the IRISA laboratory, including their computer facilities [more info]. .

5.2. Integrative Biology: (constraint-based) toolbox for network filtering

Participants: Anne Siegel [contact], Andres Aravena, Jeanne Cambefort [contact], Guillaume Collet, Damien Eveillard, Sylvain Prigent, Sven Thiele [contact].

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.

MeMap and MeMerge. We develop a workflow for the **Au**tomatic **Re**construction of **Me**tabolic networks (AuReMe). In this workflow, we use heterogeneous sources of data with identifiers from different namespaces. MeMap (**Me**tabolic network **Map**ping) consists in mapping identifiers from different namespaces to a unified namespace. Then, MeMerge (**Me**tabolic network **Merge**) merges two metabolic networks previously mapped on the same namespace. [web server].

meneco [*input*: draft metabolic network & metabolic profiles. *output*: metabolic network]. It is a qualitative approach to elaborate the biosynthetic capacities of metabolic networks. In fact, large-scale metabolic networks as well as measured datasets suffer from substantial incompleteness. Moreover, traditional formal approaches to biosynthesis require kinetic information, which is rarely available. Our approach builds upon formal systems for analyzing large-scale metabolic networks. Mapping its principles into Answer Set Programming allows us to address various biologically relevant problems [44] [27] [python package] [web server].

shogen [*input*: genome & metabolic network. *output* : functional regulatory modules]. This software is able to identify genome portions which contain a large density of genes coding for enzymes that regulate successive reactions of metabolic pathways [26] [python package].

lombarde [*input*: genome, modules & several gene-expression datasets. *output*: oriented regulation network]. This tool is useful to enhance key causalities within a regulatory transcriptional network when it is challenged by several environmental perturbations [13] [web server].

bioquali [*input*: signed regulation network & one gene-expression dataset. *output*: consistency-checking and gene-expression prediction]. It is a plugin of the Cytoscape environment. BioQuali analyses regulatory networks and expression datasets by checking a global consistency between the regulatory model and the expression data. It diagnoses a regulatory network searching for the regulations that are not consistent with the expression data, and it outputs a set of genes which predicted expression is decided in order to explain the expression inputed data. It also provides the visualization of this analysis with a friendly environment to encourage users of different disciplines to analyze their regulatory networks [6] [web server] [cytoscape plugin] .

ingranalyze [*input*: signed regulation network & one gene-expression dataset. *output*: network repair gene-expression prediction] This tool is an extension to the bioquali tool. 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] [Python package] [web server].

5.3. Dynamics: invariant-based prediction

Participants: Oumarou Abdou-Arbi, Geoffroy Andrieux, Jérémie Bourdon [contact], Jeanne Cambefort [contact], Damien Eveillard, Michel Le Borgne, Anne Siegel, Sven Thiele, Santiago Videla [contact].

We develop tools predicting some characteristics of a biological system behavior from incomplete sets of parameters or observations.

cadbiom. Based on Guarded transition semantic, this 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. [software][web server].

caspo: Cell ASP Optimizer This soft provides an easy to use software for learning Boolean logic models describing the immediate-early response of protein signaling networks. Given a network describing causal interactions, and a phospho-proteomics dataset, caspo is able to searches for optimal Boolean logic models explaining the dataset. Optimality includes both the size of the boolean network and the distance of predictions to real-data observations. It is useful to boolean networks inference, cancer research, drug discovery, and experimental design. It is used in the CellNOpt environment [1]. [python package] [web server].

nutritionAnalyzer. This tool is dedicated to the computation of allocation for an extremal flux distribution. It allows quantifying the precursor composition of each system output (AIO) and to discuss the biological relevance of a set of flux in a given metabolic network by computing the extremal values of AIO coefficients. This approach enables to discriminate diets without making any assumption on the internal behaviour of the system [15][webserver][software and doc].

POGG. The POGG software allows scoring the importance and sensibility of regulatory interactions with a biological system with respect to the observation of a time-series quantitative phenotype. This is done by solving nonlinear problems to infer and explore the family of weighted Markov chains having a relevant asymptotic behavior at the population scale. Its possible application fields are systems biology, sensitive interactions, maximal entropy models, natural language processing. It results from our collaboration with the LINA-Nantes [2][matlab package].

5.4. Sequence annotation

Participants: François Coste [contact], Aymeric Antoine-Lorquin, Catherine Belleannée [contact], Gaëlle Garet, Olivier Quenez, Jacques Nicolas.

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

We develop tools for discovery and search of complex pattern signatures within biological sequences, with a focus on protein sequences.

Logol Logol is a swiss-army-knife for Pattern matching on DNA/RNA/Protein sequences, using a high-level grammar to permit a large expressivity. Allowed patterns can consist in a combination of motifs, structures (stem-loops, repeats), indels etc. It allows pseudo-knot identification, context sensitive grammatical formalism and full genome analysis. Possible fields of application are the detection of mutated binding sites or stem-loop identification (e.g. in CRISPR [2] [10]) [software]

Protomata learner This tool is a grammatical inference framework suitable for learning the specific signature of a functional protein family from unaligned sequences by partial and local multiple alignment and automata modeling. It performs a syntactic characterization of proteins by identification of conservation blocks on sequence subsets and modelling of their succession. Possible fields of application are new members discovery or study (for instance, for site-directed mutagenesis) of, possibly non-homologous, functional families and subfamilies such as enzymatic, signaling or transporting proteins [38][4] [web server]

5.5. Integration of our tools in larger software environments

Most of our software were designed as "bricks" that can combined through workflow application such as Mobyle. It worths considering them into larger dedicated environments to benefit from the expertise of other research groups.

Web servers In collaboration with the GenOuest ressource center, most our tools are made available through several web portals.

- The **mobyle@GenOuest portal** is the generic web server of our ressource center. It hosts the ingranalysis, meneco, caspo, lombarde and shogun tools [website].
- The **Mobyle@Biotempo server** is a mobyle portal for system biology with formal approaches. It hosts the memap, memerge, meneco, ingranalysis, cadbiom and pogg tools [website].

Dr Motif This resource aims at the integration of different software commonly used in pattern discovery and matching. This resource also integrates Dyliss pattern search and discovery software [website].

ASP4biology and BioASP It is a meta-package to create a powerful environment of biological data integration and analysis in system biology, based on knowledge representation and combinatorial optimization technologies (ASP). It provides a collection of python applications which encapsulates ASP tools and several encodings making them easy to use by non-expert users out-of-the-box. [Python package] [website].

ASP encodings repository This suite comprises projects related to applications of Answer Set Programming using Potassco systems (the Potsdam Answer Set Solving Collection, bundles tools for Answer Set Programming developed at the University of Potsdam). These are usually a set of encodings possibly including auxiliary software and scripts [respository].

6. New Results

6.1. Data integration

Participants: Jacques Nicolas, Andres Aravena, Charles Bettembourg, Jérémie Bourdon, Jeanne Cambefort, Guillaume Collet, Olivier Dameron, Damien Eveillard, Julie Laniau, Sylvain Prigent, Anne Siegel, Sven Thiele, Valentin Wucher.

Metabolic network reconstruction: combinatorial gap-filling method We introduced an exhaustive gap-filling procedure on the first metabolic network for a macroalgea (Ectocarpus Siliculosus). As this species is a non benchmark model, this issue is related to hard combinatorial optimization problems. To that matter, we took advantages of the latest improvement of Answer Set Programming solvers (combination of clasp and unclasp) and introduced a new model of the network expansion problem. [*G. Collet, D. Eveillard, S. Prigent, A. Siegel, S. Thiele*] [27]

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

Identification of functional gene units in non benchmark models We introduced the concept of "shortest genome segments" (SGS) to detect functional units on exotic species, such as extremophiles, that are by nature unrefined. They correspond to genome portion which contain a large density of genes coding for enzymes which regulate successive reactions of metabolic pathways. There identification is a hard optimization combinatorial problem. We relied on the declarative modeling power of answer set programming (ASP) to encode the identification of shortest genome segments and prove that SGS are stable in (i) computational time and (ii) ability to predict functional units when one deteriorates the biological knowledge [*D. Eveillard, A. Siegel, S. Thiele*] [26]

Refinement of regulatory network from genomic, expression data and functional unit data We integrated heterogeneous information from two types of network predictions to determine a causal explanation for the observed gene co-expression. We modeled this integration as a combinatorial optimization problem. We demonstrated that this problem belongs to the NP-hard complexity class. We proposed an heuristic approach to have an approximate solution in a practical execution time. Our evaluation showed that the E.coli regulatory network resulting from the application of this method has higher accuracy than the putative one built with traditional tools. Applications to the mining bacterium *Acidithiobacillus ferrooxidans* allowed analyzing the relevance of central regulators. [*A. Aravena, D. Eveillard, A. Siegel*] [23], [13] [Thesis]

Reconstruction of a protein interaction network for archaebacteria To gain insights into genomic maintenance processes in hyperthermophilic archaea, a protein-interaction network centered on informational processes of *Pyrococcus abyssi* was generated by affinity purification coupled with mass spectrometry. We have proposed a graph theoretic analysis of this network including statistical (e.g. clusterisation coefficients) and topological aspects (bicluster analysis, search of a maximal interaction skeleton), which helps network interpretation in terms of formation of complexes or interaction dynamics. [*J. Nicolas*] [20] [Online publication]

Knowledge evolution in ontologies We studied the impact of an ontology evolution on its structural complexity. As a case study we used sixty monthly releases of the Gene Ontology and its three independent branches i.e. biological processes (BP), cellular components (CC) and molecular functions (MF). For each release, we measured complexity by computing metrics related to the size, the nodes connectivity and the hierarchical structure. We showed that the variation of the number of classes and relations in an ontology does not provide enough information about the evolution of its complexity. However, connectivity and hierarchy-related metrics revealed different patterns of values as well as of evolution for the three branches of the Gene Ontology [*O. Dameron, C. Bettembourg*] [17], [14] [Online publication] [Thesis]

Treatment process representation for breast cancer patients. The general cancer registry of Poitou-Charentes developed a multiple source information system covering diseases, anatomical structures and cytopathology. We proposed an algorithm for representing and analyzing the patient's treatment process. An expert compared the original data with our representation and computed a score of dissimilarity. The results showed that an integrated information system can successfully analyze the data to determine whether they comply with the guidelines [*O. Dameron*] [31].

AphidAtlas project We began a collaboration with the AphidAtlas project for defining the structure of an ontology of aphids anatomy and development [*O. Dameron*] [30].

6.2. Asymptotic dynamics

Participants: Anne Siegel, Oumarou Abdou-Arbi, Geoffroy Andrieux, Jérémie Bourdon, Jeanne Cambefort, Damien Eveillard, Michel Le Borgne, Vincent Picard, Sven Thiele, Santiago Videla.

Learning families of boolean signaling networks We propose the use of ASP to explore the space of feasible logic models of a signaling network. To that matter, we exhaustively enumerate the set of sub-optimal boolean logical models which are compatible with both the topology of a knowledge-based influence graph and the observed response of the system to several perturbations (phosphorylation datasets). We illustrate the importance of characterizing such a family of models in a global and exhaustive manner by revisiting a model of pro-growth and inflammatory pathways in human liver cells and studying the variability with the set of compatible models. [*A. Siegel, S. Thiele, S. Videla*] [18] [Online publication]

Control the steady-state response of qualitative signaling networks: intervention sets The minimal intervention set problem roughly consists in identifying the perturbation that can be undergone over a signaling network to predict a fixed expected behavior. We have provided a precise characterization of the minimal intervention set problem relying on three-valued logic and fixpoint semantics. We address this problem within ASP and using real-world biological benchmarks we show that it greatly outperforms previous work using dedicated algorithms. [*A. Siegel, S. Videla*] [19] [Online publication]

Reachability in dynamical signaling networks: cut sets In the scope of discrete finite-state models of interacting components, we present a novel algorithm for identifying sets of local states of components whose activity is necessary for the reachability of a given local state. Those sets are referred to as cut sets; they provide potential therapeutic targets that are proven to prevent molecules of interest to become active, up to the correctness of the model. Our method is based on the so-called Graph of Local Causality and form an under-approximation of the complete minimal cut sets of the dynamics. It makes tractable the formal analysis of very large scale networks. [*G. Andrieux, M. Le Borgne*] [28], [12] [Online publication] [Thesis]

Exploring metabolism flexibility through quantitative study of precursor sets for system outputs We extended a Flux-Balanced-Analysis approach to quantify the precursor composition of each system output and to discuss the biological relevance of a set of flux in a given metabolic network. The composition is called contribution of inputs over outputs [AIO]. In order to further investigate metabolic network flexibility, we have proposed an efficient local search algorithm computing the extremal values of AIO coefficients. This approach enables to discriminate diets without making any assumption on the internal behaviour of the system. [*J. Bourdon, O. Abdou-Arbi, A. Siegel*] [15], [11] [Thesis]

6.3. Sequence annotation

Participants: François Coste, Aymeric Antoine-Lorquin, Catherine Belleannée, Guillaume Collet, Gaëlle Garet, Clovis Galiez, Laurent Miclet, Olivier Quenez, Jacques Nicolas, Valentin Wucher.

Refinement of mi-RNA regulation network thanks to concept analysis MicroRNAs (miRNAs) are small RNA molecules that bind messenger RNAs (mRNAs) to silence their expression. To improve the discrimination between true and false interactions during their prediction, we defined a repair process based on the hypothesis that the true graph is formed by interaction modules represented by formal concepts, i.e. set of miRNAs having the same regulation profile. To validate our hypothesis and method, we have extracted parameters from a biological miRNA/mRNA network and used them to build random networks. Each repaired network can be evaluated with a score balancing the number of edge changes and the conceptual adequacy in the spirit of the minimum description length principle. [*J. Nicolas, V. Wucher*] [32]

Analogical proportions and the factorization of information in distributive lattices. We have conducted theoretical studies to elucidate whether formal concept lattices can have properties that could be used in further studies. In this direction, analogical proportions are statements involving four entities, of the form 'A is to B as C is to D'. They play an important role in analogical reasoning. They have been formalized in both a propositional logic setting and an algebraic setting. We define and study analogical proportions in the general setting of lattices, and more particularly of distributive lattices. We discussed the decomposition of analogical proportions in canonical proportions as well as the resolution of analogical proportion equations, and illustrate especially on the case of Boolean lattices, which reflects the logical modeling. [*L. Miclet*] [24], [29]

Bioinformatics and Artificial Intelligence In this book chapter, we introduce the main objects studied in Bioinformatics at different levels (the macromolecules, their interactions as well as the knowledge formalization or extraction) and present meanwhile a survey of the contribution and influence of Artificial Intelligence to this research field on related key tasks (gene prediction, functional annotation, structure prediction, transcriptomics analysis, network acquisition and analysis, knowledge integration and formalization, information retrieval and extraction from documents, ...). [*F. Coste*] [33]

Genome studies: fast assembly and SNP identification This work is a follow-up of collaborations with the GenScale team and the GenOuest platform. We reported the first identification of a set of SNPs isolated from the genome of *I. ricinus* - an important vector of pathogens in Europe, by applying a reduction of genomic

complexity, pyrosequencing and new bioinformatics tools[21] [Online publication]. We also contributed to show that the genome assembly program MINIA is successfully able to assemble a 100 Mbp genome on a very low-end, low-power system with 512 MB RAM and a 32 GB flash drive such as a Raspberry Pi. [*G. Collet, O. Quenez*] [34][Online publication]

7. Partnerships and Cooperations

7.1. Regional Initiatives

7.1.1. Regional partnership with computer science laboratories in Nantes

Participants: Anne Siegel, Jérémie Bourdon, Damien Eveillard, François Coste, Jacques Nicolas, Oumarou Abdou-Arbi, Vincent Picard, Santiago Videla, Sven Thiele.

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 process. The Ph-D students V. Picard and J. Laniau are also co-supervised with members of the LINA laboratory.

7.1.2. Regional partnership in Marine Biology

Participants: Anne Siegel, Catherine Belleannée, Jérémie Bourdon, Jeanne Cambefort, François Coste, Damien Eveillard, Jacques Nicolas, Guillaume Collet, Clovis Galiez, Gaëlle Garet, Julie Laniau, Vincent Picard, Sylvain Prigent.

A strong application domain of the Dyliss project is marine Biology. This application domain is co-developped 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).

7.1.3. Regional partnership with Inra and Health

Participants: Oumarou Abdou-Arbi, Geoffroy Andrieux, Aymeric Antoine-Lorquin, Catherine Belleannée, Charles Bettembourg, François Coste, Olivier Dameron, Michel Le Borgne, Jacques Nicolas, Anne Siegel, Valentin Wucher.

We have a strong and long term collaboration with biologists of INRA in Rennes : PEGASE and IGEEP units. This partnership is acted by the co-supervision of one post-doctorant and the co-supervision of several PhD students. The Ph-D thesis of O. Abdou-Arbi [11] and C. Bettembourg were supported by collaborations with the PEGASE laboratory [14]. This collaboration is also reinforced by collaboration within ANR contracts (Lepidolf, MirNadapt, FatInteger).

We also have a strong and long term collaboration with the IRSET laboratory at Univ. Rennes 1, acted by the defense of the co-supervised Ph-D thesis of G. Andrieux [12]. This partnership is reinforced by the ANR contract Biotempo. It was also supported in the framework of the previous CPER by a project, BasicLab, on a lab on chip for cell assays. Future studies will focus on the understanding of the metabolism of xenobiotics, funded by Anses.

7.2. National Initiatives

7.2.1. Long-term contracts

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

Participants: Anne Siegel, Jérémie Bourdon, François Coste, Damien Eveillard, Gaëlle Garet, Jacques Nicolas, Andres Aravena, Sven Thiele, Santiago Videla.

Cooperation with Univ. of Chile (MATHomics, A. Maass) on methods for the identification of biomarkers and software for biochip design. It aims at combining automatic reasoning on biological sequences and networks with probabilistic approaches to manage, explore and integrate large sets of heterogeneous omics data into networks of interactions allowing to produce biomarkers, with a main application to biomining bacteria. Co-funded by Inria and CORFO-chile from 2012 to 2022, the program includes a co-advised Ph-D student (A. Aravena) [13] and a post-doc (S. Thiele). 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 2013.

7.2.1.2. ANR Idealg

Participants: Anne Siegel, Catherine Belleannée, Jérémie Bourdon, Jeanne Cambefort, François Coste, Olivier Dameron, Damien Eveillard, Jacques Nicolas, Guillaume Collet, Clovis Galiez, Gaëlle Garet, Sylvain Prigent.

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 to workpackages 1 (establishment of a virtual platform for integrating omics studies on seaweed) and 4 (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) within workpackage 5 .[details]

7.2.2. Methodology: ANR Biotempo

Participants: Anne Siegel, Jérémie Bourdon, François Coste, Damien Eveillard, Jacques Nicolas, Michel Le Borgne, Geoffroy Andrieux, Andres Aravena, Vincent Picard, Sylvain Prigent, Santiago Videla.

The BioTempo projects aims at developing some original methods for studying biological systems. The goal is to introduce partial quantitative information either on time or on component observations to gain in the analysis and interpretation of biological data. Three biological applications are considered regulation systems used by biomining bacteria, TGF-$beta$ signaling and initiation of sea-urchin translation. It is funded by ANR Blanc (SIMI2) and coordinated by A. Siegel from 2011 to 2014. Teams involved include LINA (Nantes), I3S (Nice), DIMPP (Montpellier), Contrainte project team (Inria), IRSET (Rennes) and Station biologique de Roscoff [details]

7.2.3. Proof-of-concept on dedicated applications

7.2.3.1. ANR Fatinteger

Participants: Aymeric Antoine-Lorquin, Catherine Belleannée, Jacques Nicolas, Olivier Quenez, Anne Siegel.

This project (ANR Blanc SVE7 "biodiversité, évolution, écologie et agronomie" from 2012 to 2015) is leaded by INRA UMR1348 PEGASE (F. Gondret). Its goal is the identification of key regulators of fatty acid plasticity in two lines of pigs and chickens. To reach these objectives, this project has for ambition to test some combination of statistics, bioinformatics and phylogenetics approaches to better analyze transcriptional data of high dimension. Data and methods integration is a key issue in this context. We work on the recognition of specific common cis-regulatory elements in a set of differentially expressed genes and on the regulation network associated to fatty acid metabolism with the aim of extracting some key regulators.

7.2.3.2. ANR Lepidolf

Participants: François Coste, Jacques Nicolas.

The LEPIDOLF project aims at better understanding olfactory mechanisms in insects. The goal is to establish the antennal transcriptome of the cotton leafworm Spodoptera littoralis, a noctuid representative of crop pest insects. It is funded by ANR call Blanc and coordinated by E. Jacquin-Joly from UMR PISC (INRA Versailles) from 2009 to 2013. Our contribution is to use grammatical inference to build characteristic signatures of the Olfactory Receptor family, which will be used to scan directly 454-sequencing reads and available partial cDNAs of genes expressed in the antenna of Lepidoptera or deduced proteins.

7.2.3.3. ANR Mirnadapt

Participants: Jacques Nicolas, Catherine Belleannée, Anne Siegel, Valentin Wucher.

This ANR project is coordinated by UMR IGEPP, INRA Le Rheu (D. Tagu) and funded by ANR SVSE 6 "Génomique, génétique, bioinformatique, biologie systémique" from 2012 to 2014. This cooperation is strengthened by a co-tutored PhD thesis (V. Wucher). It proposes an integrative study between bioinformatics, genomics and mathematical modeling focused on the transcriptional basis of the plasticity of the aphid reproduction mode in response to the modification of environment. An important set of differentially expressed mRNAs and microRNAs are available for the two modes, asexual parthenogenesis and sexual reproduction. Our work is to combine prediction methods for the detection of putative microRNA/mRNA interactions as well as transcription factor binding sites from the knowledge of genomic sequences and annotations available on this and other insects. The results will be integrated within a coherent putative interaction network and serve as a filter for the design of new targeted experiments with the hope to improve functional annotations of implied genes.

7.2.3.4. ANR Pelican

Participant: François Coste.

The PELICAN project addresses competition for light in the ocean. It proposes an integrative genomic approach of the ecology, diversity and evolution of cyanobacterial pigment types in the marine environment, which arises from differences in the composition of the light-harvesting complexes (PBS). Our work is to build characteristic signatures of targeted PBS enzymes. This ANR project (génomique et biotechnologies végétales) is coordinated by F. Partensky (CRNS Roscoff) from 2010 to 2013.

7.2.4. Programs funded by research institutions

7.2.4.1. Inria Bioscience Ressource

Participants: Claudia Hériveau, Jacques Nicolas.

This project started in november 2011 and aims at promoting bioinformatics software and resources developed by Inria teams and their partners. A web portal will be deployed to allow users to test the software online. A tool is also developed to enhance the search of a specific resource using different criteria. The project is funded by Inria ADT program from 2011 to 2013, involves 8 research teams and is coordinated by the GenOuest platform and the Dyliss team (J. Nicolas and O. Collin).

7.2.4.2. PEPS VAG

Participants: François Coste, Jacques Nicolas, Clovis Galiez.

PEPS VAG started a collaboration between IMPMC UMR 7590, Institut de biologie de l'Ecole Normale Supérieure (IBENS) UMR8197, Atelier de Bioinformatique UPMC and Dyliss. It aims at defining the needs and means for a larger project about viruses in marine ecosystems. Indeed, we aim at developing new methods based on both sequential and structural information of proteins to improve the detection of viral sequences in marine metagenomes, to identify new viruses and to compare the viral populations specifically associated with different environment parameters (temperature, acidity, nutriments...) and ultimately to connect them with the potential hosts identified by population sequencing.

7.3. European Initiatives

7.3.1. Collaborations with Major European Organizations

Partner: EBI (Great-Britain)

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

Partner: Potsdam university (Germany)

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

7.4. International Initiatives

7.4.1. Inria Associate Teams

7.4.1.1. INTEGRATIVEBIOCHILE

Title: Bioinformatics and mathematical methods for heterogeneous omics data

Inria principal investigator: Anne Siegel

International Partner (Institution - Laboratory - Researcher):

University of Chile (Chile) - Center for Mathematical Modeling - Alejandro Maass

Duration: 2011 - 2013

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

IntegrativeBioChile is an Associate Team between Inria project-team "Dyliss" and the "Laboratory of Bioinformatics and Mathematics of the Genome" hosted at CMM at University of Chile. The Associated team is funded from 2011 to 2013. The project aims at developing bioinformatics and mathematical methods for heterogeneous omics data. Within this program, we funded long and short stay visitings in France.

7.4.2. 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 and the co-supervised of A. Aravena contribute to reinforce the complementarity of both Chilean and French teams. In 2013, a workshop was organized in Chile to develop new French-Chilean collaboration within the framework of the CIRIC center. See Sec. 7.2.1 for details.

7.4.3. Participation In other International Programs

7.4.3.1. Argentina - MinCYT-Inria 2011-13

Partner: Universidad Nacional de Cordoba, *Grupo de Procesamiento de Lenguaje Natural (PLN)*, Argentina.

Title: Modélisation linguistique de séquences génomiques par apprentissage de grammaires

Financial support: MinCYT-Inria program 2011-13

The projects aims at developing new grammatical inference methods to learn automatically linguistic models of genomic sequences.

7.4.3.2. International joint supervision of PhD

Title: Introduction des approches combinatoires dans des modèles probabilistes pour la découverte d'évènements de régulation d'un système biologique à partir de données hétérogènes [*A. Aravena*]

Inira principal investigator: Anne Siegel

International Partners (Institution - Laboratory - Researcher):

University of Chile (Chile) - Center of Mathematical Modelling - Alejandro Maass

Duration: Jul 2011 - Dec 2013

Title: Analyse automatisée et générique de réseaux métaboliques en nutrition [*O. Abdou-Arbi*]

Inria principal investigator: Anne Siegel

International Partner (Institution - Laboratory - Researcher):

University of Ouagadougou (Burkina Faso) - Department of mathematics - T. Tabsoba.

Duration: October 2010 - September 2013

Title: Applying logic programming to the construction of robust predictive and multi-scale models of bioleaching bacteria [*S. Videla*]

Inria principal investigator: Anne Siegel

International Partner (Institution - Laboratory - Researcher):

University of Postdam (Germany). Department of computer science. T. Schaub.

Duration: October 2011 - September 2014

7.5. International Research Visitors

7.5.1. Visits of International Scientists

- **Germany**. Department of Computer Science, Potsdam [T. Schaub]
- **Chile**. Centro de Modelimiento Matematico, Santiago [A. Maass, N. Loirà]
- **Burkina-Faso**. Laboratoire de mathématiques, Ouagadougou [T. Tabsoba]

7.5.1.1. Internships

Andres Aravena

Subject: Programmation par Ensemble-Réponse pour l'identification de régulateur clés en biologie des systèmes

Date: from Jan 2013 until Jul 2013

Institution: University of Chile (Chile)

7.5.2. Visits to International Teams

- **Burkina-Faso**. Department of Computer Science, Ouagadougou. *Multi-objective methods for the static analysis of metabolic network*. Jan. 2013 (1 month) [O. Abdou-Arbi]
- **Niger**. University of Maradi. *Multi-objective methods for the static analysis of metabolic network*. Feb. 2013 (1 month) [O. Abdou-Arbi]
- **UK** EMBL-European Bioinformatics Institute. *Learning logical rules for protein signaling networks*. Feb. 2013 (2 days) [A. Siegel, S. Thiele, S. Videla]
- **UK** Brunel University *Learning logical rules for protein signaling networks*. Feb. 2013 (3 days) [A. Siegel, S. Thiele, S. Videla]
- **Germany**. Max Planck Institute (Klamt lab), Magdeburg. *Application of ASP to the control of signaling networks*. June 2013 (2 days) [S Thiele, S. Videla]

- **Argentina**. Departamento Universitario de Informática, Cordoba. *Collaboration on grammatical inference*. Jul. 2013 (14 days) [F. Coste]
- **Argentina**. Departamento Universitario de Informática, Cordoba. *Collaboration on grammatical inference*. Jul. 2013 (1 month) [G. Garet]
- **Germany**. Department of Computer Science, Potsdam. *Application of ASP to biology, meeting with Klamt and Schaub labs*. Oct 2013 (3 days) [A. Siegel, S. Videla]
- **Germany**. Department of Computer Science, Potsdam. *Application of ASP for sequence annotation*. Oct. Nov. Dec. 2013 (3 months) [G. Garet]
- **Chile**. Centro de Modelimiento Matematico, Santiago. *Applications of ASP*. Nov. & Dec 2013 (2 monthes) [S. Videla]
- **Chile**. Centro de Modelimiento Matematico, Santiago. *Metabolic modeling of bacteria*. Dec. 2013 (14 days) [D. Eveillard]
- **Chile**. Centro de Modelimiento Matematico, Santiago. *Data integration*. Dec. 2013 (7 days) [A. Siegel, S. Prigent, J. Laniau, V. Picard, F. Coste]

8. Dissemination

8.1. Scientific Animation

8.1.1. Administrative functions: scientific committees, journal boards

- Scientific Advisory Board of GDR BIM " Molecular Bioinformatics"[J. Nicolas].
- Member of the IRISA laboratory council [F. Coste]
- Member of the Inria Rennes center council [A. Siegel]
- Scientific Advisory Board of Biogenouest [J. Bourdon, J. Nicolas, A. Siegel].
- Expertise for AERES (LIFL, LRI) [J. Nicolas, A. Siegel] & ANR (programme blanc) [A. Siegel]
- Academic editor: Plos One [J. Bourdon]
- Ecole Normale recruitment jury [A. Siegel]
- Recruitment committees: junior researcher (Inra) [J. Nicolas], assistant professor (Univ. Rennes) [A. Siegel], assistant professor (Univ. Montpellier) [O. Dameron].
- Reviewer: Advances in Mathematics [A. Siegel], Algorithms for Molecular Biology [F. Coste], American Medical Informatics Association conference [O. Dameron], Bioinformatics [O. Dameron], BMC Bioinformatics [O. Dameron, A. Siegel], Briefings in Bioinformatics [O. Dameron], Bulletin of the Belgium Math. Society [A. Siegel], Computers in Biology and Medicine [F. Coste], Journal of Biomedical Semantics [O. Dameron], Machine Learning [F. Coste], Semantic Web Applications and Tools for Life Sciences workshop [O. Dameron].
- Member of SCAS (Service Commun d'Action Sociale) of Univ. Rennes 1 [C. Belleannée]

8.1.2. Meetings

- **Workshop on Integrative-Omics** A workshop was organized in Pucon, south Chile, in December 2013. The workshop gathered about 40 scientists from Chile, France and Germany. Nine invited lectures were given about the modeling of biological systems and data with various computational approaches (MILP, ASP, graph complexity, probabilistic and stochastic approaches). 15 young scientist talks were also given in complement to the lectures [website].
- **Seminar** A weekly seminar of bioinformatics is organized within the laboratory. Attendees are member of the symbiose team, biologists from Brittany and computer scientists from the laboratory. [website].

8.1.3. Conference program committees

- JOBIM [F. Coste]
- CIBB Computational Intelligence Methods for Bioinformatics and Biostatistics & PRIB International Conference on Pattern Recognition in Bioinformatics [A. Siegel]
- Program committee of Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2013) [O. Dameron].

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

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

Licence: C. Belleannée, Architecture des ordinateurs, 50h, L3 informatique, Rennes1 France.

Licence: C. Belleannée, Bases de données, 21h, L3 Miage par alternance, Rennes1 France .

Licence: G. Andrieux, TIC : Technologies de l'information et de la communication, 32h, L1, Univ. Rennes 1, France.

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

Licence: O. Dameron, C2i niveau 2, 2.5h, Univ. Rennes 1, France.

Licence: V. Picard, Scheme 14h, L1, INSA Rennes, France

Licence: V. Picard, Architecture et systèmes, 24h, L3, ENS Rennes/Univ. Rennes 1, France

Licence: V. Picard, Initiation Unix, 2h, L3, ENS Rennes, France

Licence: S. Prigent, learning PHP/SQL, 12h, L3 (3ème année ingénieur), Ensai, Rennes, France

Licence: S. Prigent, Database, 42h, L1, Ensai, Rennes, France

Licence: S. Prigent, An introduction to R, 9h, L1, Ensai, Rennes, France

Licence: V. Wucher, Introduction aux biostatistiques, 8h, L3 biologie, Rennes 1, France

Master: C. Belleannée, Préférences Logique et contraintes, 32h, M1 informatique, Rennes1 France

Master: C. Belleannée, Architecture matérielle et interface au système, 28h, M2 informatique, 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, gestion de projets en informatique, 49h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, principes de programmation et d'algorithmique, 64h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, initiation systèmes et réseaux, 4h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, modélisation des connaissances et bio-ontologies, 36h, M2 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, Bases de mathématiques, probabilités et statistiques, 65h, M1 santé publique, Univ. Rennes 1, France.

Master: O. Dameron, E-santé et réseaux hospitaliers, 7h, ESIR3, Univ. Rennes 1, France.

Master: C. Galiez, Compilation, 32h, M1 informatique, Rennes1 France

Master: G. Collet, Langage C, 8h, M1 Informatique, Univ. Rennes 1, France

Master: G. Collet, Analyse et Conception Objet, 14h, M1 MIAGE, Univ. Rennes 1, France

Master: V. Picard, Préparation à l'agrégation de mathématiques option D: épreuve de modélisation, 20h, L2, ENS Rennes/Univ. Rennes 1, France

Master: V. Picard, Aspects probabilistes en biologie des systèmes, 4h, M2, ENS Rennes/Univ. Rennes 1, France

Doctorat: A. Siegel, Programmation par ensemble-réponse (ASP) et application à la reconstruction et la correction de réseaux biologiques, 3h, Ecole thématique CNRS, Modélisation Formelle de Réseaux de Régulation Biologique, France

Doctorat: S. Videla, Answer Set Programming for Systems Biology, 10h, graduate course, Universidad de Chile, Chile.

8.2.2. Seminars

- A. Siegel. *Modeling the quantitative behaviors of biological systems*. IBISC laboratory. Evry (Jan. 2013).
- A. Siegel. *Extracting robust information from the confrontation of knowledge and observations on a biological system*. EBI, UK (Feb. 2013).
- S. Videla, *Learning Logic Models of Protein Signaling Networks with Answer Set Programming*. Brunel Univ (Feb. 2013).
- S. Videla. *(Boolean) Logic Models of Signal transduction Networks with Answer Set Programming*. Universidad de Chile (Apr. 2013).
- A. Siegel. *Extracting robust information from knowledge and observations on a biological system: a formal system framework.*. CIBB Computational Intelligence Methods for Bioinformatics and Biostatistics & PRIB International Conference on Pattern Recognition in Bioinformatics, keynote lecture. (Jun. 2013).

8.2.2.1. Internships

- Internship, from Jun. until Sep. 2013. Co-supervised by J. Nicolas, G. Garet. Student :Liantsoa Rasata Manantena. Subject: Characterization of enzyme family sequences : sulfatases
- Internship, from Jan. until Jun. 2013. Supervised by C. Belleannée. Student: Aymeric Antoine Lorquin. Subject: Identification in silico du site de fixation de la protéine CELF1, au moyen de pattern matching spécialisé.
- Internship, from Ap. until Jun. 2013. Supervised by O. Dameron. Student: Ayite Kougbeadjo. Subject: Analyse de réactions candidates pour la reconstruction de bases métaboliques basée sur les connaissances symboliques chez Ectocarpus Siliculosus
- Internship, from Apr. until Jun. 2013. Supervised by J. Nicolas and V. Wucher. Student: Lucas Le Lann. Subject: Analyse exploratoire d'un réseau d'interactions par extraction de composants analogues appliqué au puceron du pois..
- Internship, from Apr. until Jun. 2013. Supervised by G. Andrieux. Student: Jean Coquet. Subject: Modélisation du réseau d'activation du TGF-β

8.2.3. Supervision

PhD : Oumarou Abdou-Arbi *Etude de la variabilité des contributions de nutriments à un réseau métabolique : modélisation, optimisation et application en nutrition*, 30 Sept. 2013, supervised by A. Siegel and T. Tabsoba (Burkina-Faso) [11].

PhD : Geoffroy Andrieux, *Discrete approach modeling of biological signaling pathway*, 18 Jul. 2013, supervised by N. Théret (Inserm) and M. Le Borgne [12]

PhD : Andres Aravena, *Probabilistic and constraint based modelling to determine regulation events from heterogeneous biological data*, 13 Dec. 2013, supervised by A. Maass (CMM, University of Chile) and A. Siegel [13].

PhD : Charles Bettembourg, *Modélisation Méthodes sémantiques pour la comparaison inter-espèces de voies métaboliques : application au métabolisme des lipides chez l'humain, la souris et la poule*, 16 Dec. 2013, supervised by O. Dameron and C. Diot (Inra) [14].

PhD in progress : Aymeric Antoine-Lorquin, *Modèles grammaticaux au service de l'identification de marqueurs de régulation génétique dans les séquences biologiques*, started in Oct. 2013, supervised by C. Belleannée

PhD in progress : Gaëlle Garet, *Discovery of enzymatic functions in the framework of formal languages*, started in Oct. 2011, supervised by J. Nicolas and F. Coste.

PhD in progress : Clovis Galiez, *Syntactic modelling of protein structure.*, started in Oct. 2012, supervised by F. Coste.

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 : Vincent Picard, *Analyse dynamique d'algorithmes et dynamique symbolique pour l'étude de modèles semi-quantitatifs en biologie des systèmes*, started in Sept. 2012, supervised by A. Siegel and J. Bourdon.

PhD in progress : Sylvain Prigent, *Modélisation par contraintes pour le contrôle génomique et physiologique de l'adaptation des algues brunes à la salinité de l'eau*, started in Oct. 2011, supervised by A. Siegel and T. Tonon (UMR 7150, station biologique de Roscoff)

PhD in progress : Santiago Videla, *Applying logic programming to the construction of robust predictive and multi-scale models of bioleaching bacteria*, started in Nov. 2011, supervised by A. Siegel and T. Schaub (Potsdam univ).

PhD in progress : Valentin Wucher, *Modélisation d'un réseau de régulation d'ARN pour prédire des fonctions de gènes impliqués dans le mode de reproduction du puceron du pois*, started in Nov. 2011, supervised by J. Nicolas and D. Tagu (Inra)

8.2.4. Juries

- *Member of Ph-D thesis jury*. F. Nguema Ndong, univ. Poitiers [A. Siegel, présidente]. T. Jolivet, Univ. Paris Diderot [A. Siegel].

8.3. Popularization

- *Bioinfo-fr.net* Bioinfo-fr.net is a french web site where researchers, engineers and students talks about bioinformatics. We have written 6 articles for this web site on diverse subjects: metabolic networks, genome assembly, phylogenetics. [G. Collet, S. Prigent]. [more info].
- *Participation at Rennes* Village des Sciences *for French National Science day (*Fête de la Science*)*. Title : *Crazy random walks*. Description : Popularization Festival where researchers present scientific themes. [V. Picard]
- *Fête de la science (LINA, Nantes)* During the 24th of october, 250 students discovered bioinformatics, genome assembly and algorithmic by practicing games and tutorials made by our team with the help of Julien Gras and Marko Budinich (LINA) [J. Bourdon, D. Eveillard, G. Collet].
- *Organization of Sciences en Cour[t]s*. Popularization Festival where PhD students explain their thesis via short films. [S. Prigent, C. Bettembourg, G. Garet] [more info].

9. Bibliography

Major publications by the team in recent years

[1] 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://hal.inria.fr/inria-00538133

[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 CORGUILLÉ, C. SIX, M. RATIN, O. COLLIN, W. M. SCHLUCHTER, F. PARTENSKY. *CyanoLyase: a database of phycobilin lyase sequences, motifs and functions*, in "Nucleic Acids Research", November 2012, 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] 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

[9] 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

[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] O. ABDOU-ARBI. , *Etude de la variabilité des contributions de nutriments à un réseau métabolique : modélisation, optimisation et application en nutrition*, Université Rennes 1, September 2013, http://hal.inria.fr/tel-00924200

[12] G. ANDRIEUX. , *Modélisation dynamique de la signalisation cellulaire : aspects différentiels et discrets; application à la signalisation du facteur de croissance TGF-beta dans le cancer*, Université Rennes 1, July 2013, http://hal.inria.fr/tel-00926487

[13] A. ARAVENA. , *Probabilistic and constraint based modelling to determine regulation events from heterogeneous biological data*, Université Rennes 1, December 2013, http://hal.inria.fr/tel-00922346

[14] C. BETTEMBOURG. , *Méthodes sémantiques pour la comparaison inter-espèces de voies métaboliques : application au métabolisme des lipides chez l'humain, la souris et la poule*, Université Rennes 1, December 2013, http://hal.inria.fr/tel-00926498

Articles in International Peer-Reviewed Journals

[15] O. ABDOU-ARBI, S. LEMOSQUET, 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, 8 p. [*DOI :* 10.1186/1752-0509-8-8], http://hal.inria.fr/hal-00924253

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

[17] O. DAMERON, C. BETTEMBOURG, N. LE MEUR. *Measuring the evolution of ontology complexity: the gene ontology case study*, in "PLoS ONE", 2013, vol. 8, n^{o} 10 [*DOI :* 10.1371/JOURNAL.PONE.0075993], http://hal.inria.fr/hal-00877378

[18] 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

[19] 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

[20] P.-F. PLUCHON, T. FOUQUEAU, C. CREZÉ, 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 [*DOI :* 10.1371/JOURNAL.PONE.0079707], http://hal.inria.fr/hal-00911795

[21] E. QUILLERY, O. QUENEZ, P. PETERLONGO, O. PLANTARD. *Development of genomic resources for the tick Ixodes ricinus: isolation and characterization of Single Nucleotide Polymorphisms*, in "Molecular Ecology Ressources", October 2013, epub ahead of print [*DOI :* 10.1111/1755-0998.12179], http://hal.inria.fr/hal-00880072

Invited Conferences

[22] A. SIEGEL. *Extracting robust information from knowledge and observations on a biological system: a formal system framework*, in "PRIB - Eighth IAPR International Conference on Pattern Recognition in Bioinformatics - 2013", Nice, France, June 2013, http://hal.inria.fr/hal-00853839

International Conferences with Proceedings

[23] 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, Lecture Notes in Computer Science, Springer, 2014, vol. 8318, pp. 322-336 [*DOI :* 10.1007/978-3-642-54013-4_18], http://hal.inria.fr/hal-00926477

[24] N. BARBOT, L. MICLET, H. PRADE. *Analogical proportions and the factorization of information in distributive lattices*, in "10th International Conference on Concept Lattices and Their Applications (CLA)", La Rochelle, France, October 2013, http://hal.inria.fr/hal-00908005

[25] V. BERTHE, J. BOURDON, T. JOLIVET, A. SIEGEL. *Generating discrete planes with substitutions*, in "WORDS - 9th International Conference on Words - 2013", Turku, Finland, September 2013, http://hal.inria.fr/hal-00853758

[26] 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

[27] 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

[28] L. PAULEVÉ, G. ANDRIEUX, H. KOEPPL. *Under-approximating Cut Sets for Reachability in Large Scale Automata Networks*, in "25th International Conference, CAV 2013, Proceedings", Saint Petersburg, Russian Federation, Lecture Notes in Computer Science, July 2013, vol. 8044, pp. 69-84 [*DOI :* 10.1007/978-3-642-39799-8_4], http://hal.inria.fr/hal-00769447

National Conferences with Proceedings

[29] N. BARBOT, L. MICLET, H. PRADE. *Proportions analogiques et factorisation de l'information dans les treillis distributifs*, in "Journées d'Intelligence Artificielle Fondamentale (JIAF)", Aix en Provence, France, June 2013, http://hal.inria.fr/hal-00908020

[30] A. BRETAUDEAU, O. DAMERON, F. LEGEAI, Y. RAHBÉ. *AphidAtlas : avancées récentes*, in "Proceedings of BAPOA 2013 MOP. INRA, CIRAD Lavalette Campus Montpellier, France", Montpellier, France, M. UZEST (editor), 2013, [In French], http://hal.inria.fr/hal-00913155

[31] A. ROLLET, G. DEFOSSEZ, O. DAMERON, CORIM, POITOU-CHARENTES, CRISAP, POITOU-CHARENTES, P. INGRAND. *Développement et évaluation d'un algorithme de représentation des parcours de soins de patientes atteintes d'un cancer du sein à partir des données d'un système d'information régional*, in "Proceedings of the conference Évaluation Management Organisation Information Santé (EMOIS2013)", Nancy, France, 2013, [In French, short abstract], http://hal.inria.fr/hal-00913172

Conferences without Proceedings

[32] V. WUCHER, D. TAGU, J. NICOLAS. *Edge Selection in a Noisy Graph by Concept Analysis: Application to a Genomic Network*, in "ECDA - European Conference on Data Analysis - 2013", Luxembourg, Luxembourg, July 2013, http://hal.inria.fr/hal-00924413

Scientific Books (or Scientific Book chapters)

[33] F. COSTE, C. NÉDELLEC, T. SCHIEX, J.-P. VERT. *Bioinformatique*, in "Panorama de l'intelligence artificielle Ses bases méthodologiques, ses développements", P. MARQUIS, O. PAPINI, H. PRADE (editors), Cépaduès, June 2013, vol. 3, http://hal.inria.fr/hal-00857379

Other Publications

[34] G. COLLET, G. RIZK, R. CHIKHI, D. LAVENIER. *MINIA on a Raspberry Pi, Assembling a 100 Mbp Genome on a Credit Card Sized Computer*, in "JOBIM - Journées Ouvertes en Biologie, Informatique et Mathématiques", Toulouse, France, July 2013, http://hal.inria.fr/hal-00842027

References in notes

[35] G. ANDRIEUX, L. FATTET, M. LE BORGNE, R. RIMOKH, N. THÉRET. *Dynamic regulation of Tgf-B signaling by Tif1γ: a computational approach*, in "PLoS ONE", 2012, vol. 7, n^{o} 3 [*DOI :* 10.1371/JOURNAL.PONE.0033761], http://hal.inria.fr/hal-00851483

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

[37] 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^{o} 3, pp. 257-64

[38] S. BRADFORD, F. COSTE, M. VAN ZAANEN. *Progressing the state-of-the-art in grammatical inference by competition*, in "AI Communications", 2005, vol. 18, n^{o} 2, pp. 93-115

[39] 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

[40] P. FLAJOLET, R. SEDGEWICK. , *Analytic Combinatorics*, Cambridge University Press, 2009

[41] 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

[42] 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

[43] 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

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

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

[46] 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

[47] 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

[48] T. TONON, D. EVEILLARD, S. PRIGENT, J. BOURDON, P. POTIN, C. BOYEN, A. SIEGEL. *Toward systems biology in brown algae to explore acclimation and adaptation to the shore environment*, in "Omics : a journal of integrative biology", 2011, vol. 15, n^{o} 12, pp. 883–892