IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Activity Report 2014

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 - Inria (1/2 delegation)]
Laurent Miclet [Univ. Rennes I, Emerite Professor, HdR]

Engineers

Charles Bettembourg [CNRS, post-doc, Mirnadapt project, from Aug 2014]
Jeanne Cambefort [CNRS, permanent engineer]
Marie Chevallier [Inria, engineer, from Oct 2014]
Guillaume Collet [CNRS, post-doc, Idealg project]
Yann Guitton [CNRS, post-doc, Idealg project, from Sep 2014]

PhD Students

Aymeric Antoine-Lorquin [Univ. Rennes I]
Jean Coquet [Univ. Rennes I]
Victorien Delannée [Univ. Rennes I - Région/Anses]
Clovis Galiez [Inria]
Gaëlle Garet [Inria]
Julie Laniau [Inria]
Vincent Picard [Univ. Rennes I/ENS Rennes]
Sylvain Prigent [Univ. Rennes I, until Nov 2014]
Santiago Videla [CNRS, until Aug 2014]
Valentin Wucher [INRA, until Oct 2014]

Visiting Scientists

Philippe Bordron [Chilean post-doc, from May 2014 until Jun 2014]
Maria Paz Cortes Burgos [Chilean ph-D, from Sep 2014 until Oct 2014]
Mario Weitzer [Austrian ph-D, from Sep 2014 until Oct 2014]
Meriem Zekri [Tunisian ph-D, from Sep 2014 until Oct 2014]

Administrative Assistants

Marie-Noëlle Georgeault [Inria, until Feb. 2014]
Isobelle Kelly [Univ. Rennes I, from Feb 2014 until Dec 2014]
Marie Le Roic [Univ. Rennes I, from Dec 2014]

Others

Jérémie Bourdon [Univ. Nantes, Dyliss associate member]
Damien Eveillard [Univ. Nantes, Dyliss associate member]
Nathalie Théret [INSERM, Dyliss associate member, HdR]

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).

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) [5].

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** [45], [53] 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 [46], and proposed robust corrections [4]. 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 [1]. 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 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 [52]. 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

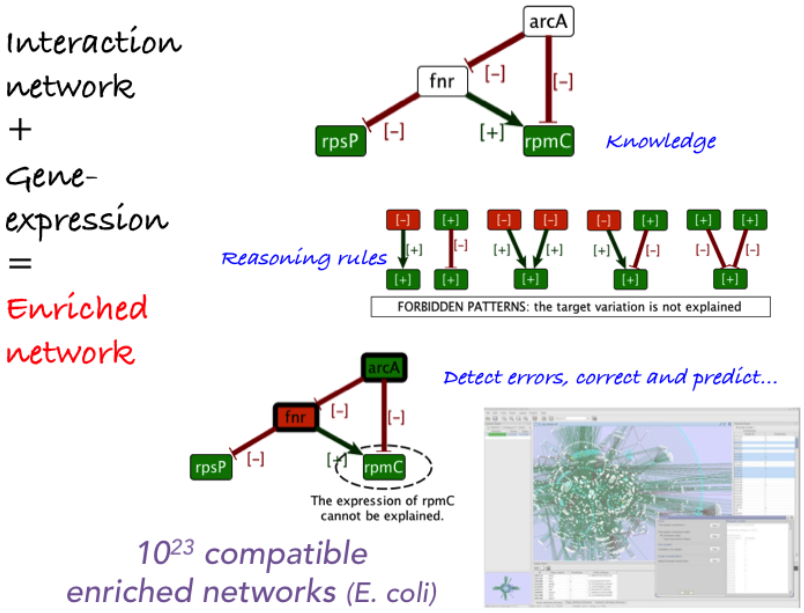

Figure 1.

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. [5], [4] ***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). .*

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.

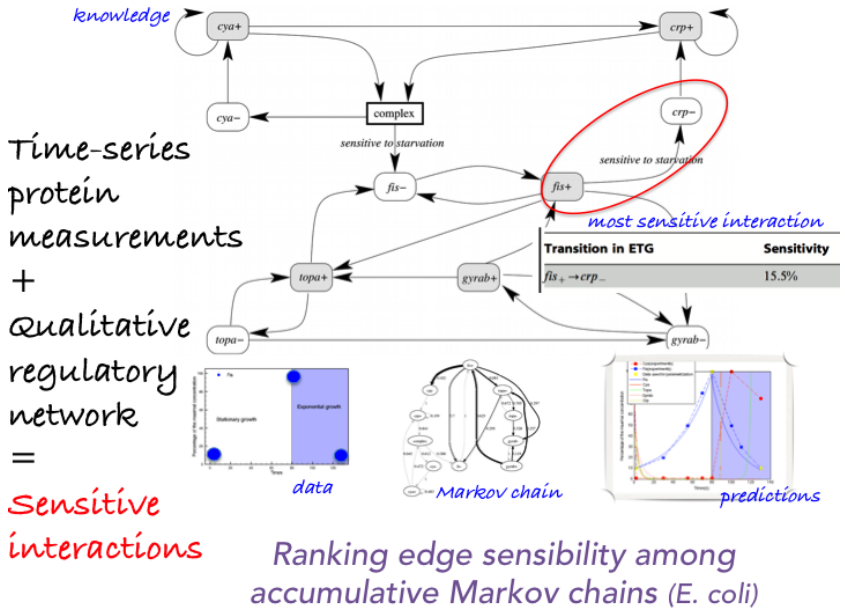

Figure 2.

Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems and identification of key interactions [1].

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 the 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 [3], leading to a tool named Protomata-learner. As an example, this tools allows us to properly model the 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 [2]. Our future goal is to further demonstrate the relevance of formal language modelling 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.

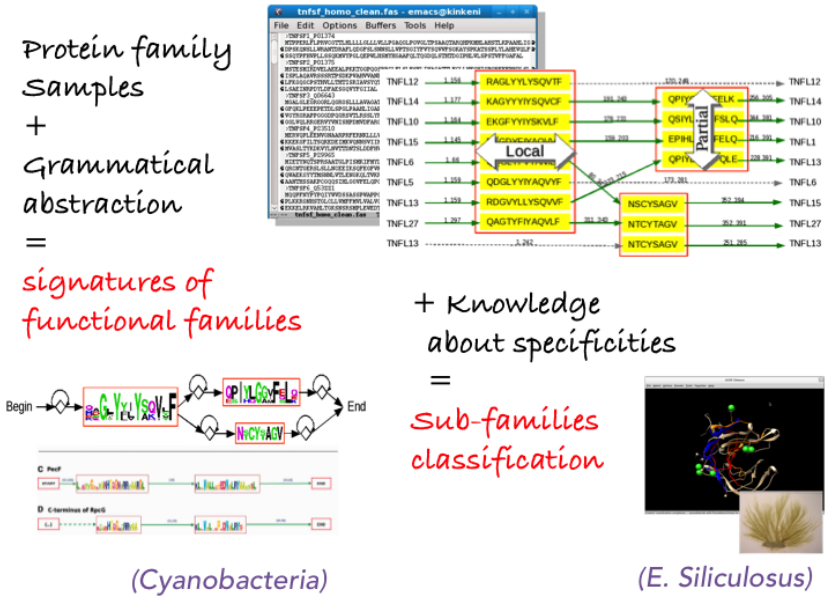

Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences, a partial local alignment is computed and an automaton is inferred, which can be considered as a signature of the family of proteins. This allows searching for new members of the family [2]. Adding further information 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 [54]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [59], [58]. 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 [7]. This tool was used for the recognition of transposable elements in *Arabidopsis thaliana* [60] or for the design of a CRISPR database [9]. 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.

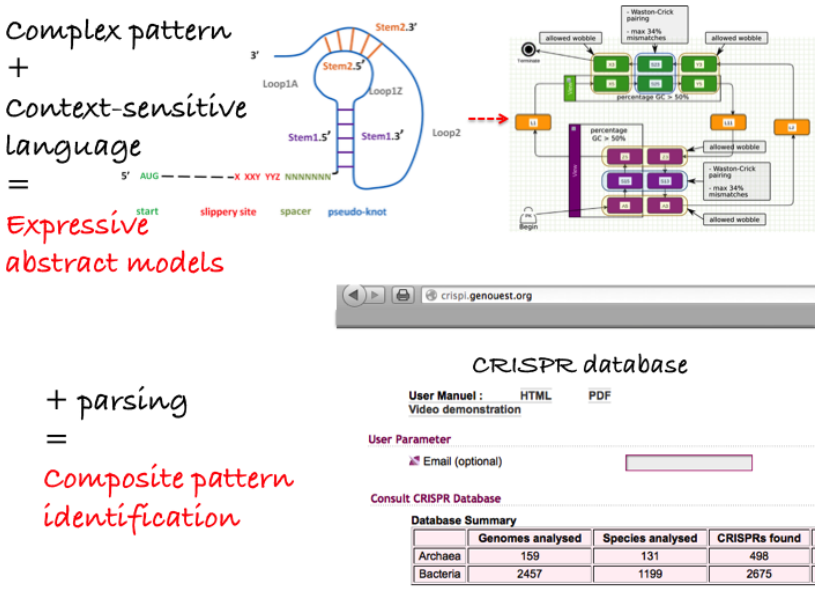

Figure 4. Graphical modeling of a pseudo-knot (RNA structure) based on the expressivity of String Variable Grammars used in the Logol framework. Combined with parsers, this leads to composite pattern identification such as CRISPR [56].

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 [22], [19]. 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 [28]. The same tool also allowed us to gain a better understanding of cyanobacteria proteins [2]. Finally, in dynamical systems, we use asymptotic analysis (tool *pogg*) to decipher the initiation of sea urchin translation [47]. 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 [5]. In parallel, in collaboration with Ifremer (Brest), we have conducted similar work to decipher protein-protein interactions within archebacteria [55]. Our sequence analysis tool (*logol*) allowed us to build and maintain a very expressive CRISPR database [9] [27].

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 [14]. The pattern matching tool *logol* also allows for a fine identification of transcription factor motifs [51] [27]. Constraints-based programming also allows us to decipher regulators

of reproduction for pea aphids [30], [13]. Semantic-based analysis was useful for interpreting differences of gene expression in pork meat [20].

We are less involved in **bio-medical applications** 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 [46]. 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 [23] and N. Theret group at Inserm/Irset (Rennes) [15]. 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. New Software and Platforms

5.1. Platforms and toolboxes

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

Most of our 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].

We present here three toolboxes which each contain complementary tools with respect to their targeted sub-domain of bioinformatics.

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

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 [57] [50] [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 [48] [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 [26] [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 [5] [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 [4] [Python package] [web server].

Unifier. [*input*: sbml file with Palsson's metabolites identifiers *output*: sbml file with standard identifiers for metabolites]. This software is a Decision Support Tool to help biologists to normalize a file, containing Palsson's identifiers to refer to reactions and metabolites, using well known identifiers. Submit a list of Palsson identifiers to retrieve the corresponding database entries. Typically it maps with Metacyc identifiers but it would be used with Kegg or other databases later. A Unifier web service will be soon available.

NetWikiMaker. This tool generates (half) automatically a wiki on our reconstruction workflow. It contains information and data about the network reconstruction process such as different versions of draft metabolic networks files, parameters of tools, log files. It also displays the reactions, genes and metabolites that the workflow has found to be involved in the metabolic network, and provides a powerful search tool.

5.1.2. Dynamics and invariant-based prediction

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 [14][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 [1][matlab package].

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

5.1.3. Sequence annotation

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] [9]) [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 [49][3] [web server]

5.1.4. Integration of toolboxes and platforms in webservices

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].
- The **GenOuest galaxy portal** now provides access to most tools for integrative biology and sequence annotation (access on demand).

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].

5.2. New tools for integrative biology

Participants: Anne Siegel [contact], Jeanne Cambefort [contact], Guillaume Collet, Damien Eveillard, Sylvain Prigent, Marie Chevallier.

The tools MeMap and MeMerge were complemented with new tools in order to analyze reference networks from litterature database and to vizualize the product of reconstructed metabolic networks.

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

Unifier. [*input*: SBML file with Palsson's metabolites identifiers *output*: sbml file with standard identifiers for metabolites]. This software is a Decision Support Tool to help biologists to normalize a file, containing Palsson's identifiers to refer to reactions and metabolites, using well known identifiers. Submit a list of Palsson identifiers to retrieve the corresponding database entries. Typically it maps with Metacyc identifiers but it would be used with Kegg or other databases later. A Unifier web service will be soon available.

NetWikiMaker. This tool generates (half) automatically a wiki on our reconstruction workflow. It contains information and data about the network reconstruction process such as different versions of draft metabolic networks files, parameters of tools, log files. It also displays the reactions, genes and metabolites that the workflow has found to be involved in the metabolic network, and provides a powerful search tool.

5.3. New tools for dynamics

Participants: Jérémie Bourdon [contact], Jeanne Cambefort [contact], Damien Eveillard, Anne Siegel, Nathalie Théret, Santiago Videla [contact].

In 2014, the tool caspo was extended to new functionnalities.

caspo: Cell ASP Optimizer In the new version of caspo, *automated inference* of logical networks from experimental data allows for identifying admissible large-scale logic models saving a lot of efforts and without any a priori bias. Next, once a family a logical networks has been identified, one can suggest or *design new experiments* in order to reduce the uncertainty provided by this family. Finally, one can look for *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. Altogether, this constitutes a pipeline for automated reasoning on logical signaling networks. Hence, the aim of caspo is to implement such a pipeline providing a powerful and easy-to-use software tool for systems biologists. [doc and download as a python package] [web server].

6. New Results

6.1. Highlights of the Year

Four PhD theses were defended this year. They evidenced that ASP-technologies are now mature enough to perform data integration of large-scale bio-molecular datasets: classification of families of proteins [10], reconstruction of regulatory networks [13], reconstruction of metabolic network [11], and modelling of the discrete dynamics of a signalling or a regulatory network [12]. Importantly, symbolic classification technics have been adapted to exhibit relevant biological features: we used both formal concept analysis and semantic-based analysis for sequence and network analysis.

6.2. Data integration

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

Pan-genomic metabolic network of *Ectocarpus siliculosus* : We introduced the first metabolic network for the non-classical species *E. Siliculosus*, called EctoGEM. The reconstruction process includes draft reconstruction based on sequence and functional annotation analysis. It is followed by a combinatorial gap-filling process using the Meneco software based on answer set programming, a semantic analysis of the completion and a manual curation. This reconstruction enables a better understanding of organism biology and a reannotation of its genome. [*J. Cambefort, G. Collet, O. Dameron, D. Eveillard, S. Prigent, A. Siegel*] [22], [11]

New insights on bacteria associated with brown algae As an application of our tools for the reconstruction of metabolic networks, we have contributed to the analysis of the genome of a bacteria which lives in symbiosis with brown algae by investigating candidates for metabolic exchanges between the bacteria and the algae. [*G. Collet, J. Cambefort, A. Siegel*] [19] [Online publication]

Modeling parsimonious putative regulatory networks 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. [*A. Aravena, A. Siegel, D. Eveillard*] [26] [Online publication]

Modeling of a gene network between mRNAs and miRNAs to predict gene functions involved in phenotypic plasticity in the pea aphid and non coding RNA in pea aphid During its PhD, V. Wucher has built the first network model of gene regulation by microRNAs in pea aphid. The thesis has studied the discrimination between embryos development towards either sexual or asexual reproduction types in the pea aphid *Acyrthosiphon pisum*, at the genomic level. The study of the post-transcriptional regulation network implies both the identification of regulated elements during embryogenesis and the identification of the interaction modules between microRNAs and mRNAs using formal concept analysis. It helps the understanding of regulation differences between sexual and asexual embryogenesis. Moreover, it is the first step towards the modeling of the entire set of genes regulations in pea aphid during embryogenesis. [*V. Wucher, J. Nicolas, F. Legeai (Genscale team)*] [13], [39], [30]

Using a large-scale knowledge database on reactions and regulations to exhibit key regulators A new formalism of regulated reactions combining biochemical transformations and regulatory effects was proposed to unify the different mechanisms contained in knowledge libraries. Based on a related causality graph, an algorithm was developed to propose a reasonable set of upstream regulators from lists of target molecules. Scores were added to candidates according to their ability to explain the greatest number of targets or only few specific ones. The method was validated on a real example related to glycolysis. [*P. Blavy, A. Siegel*] [18] [Online publication]

Semantic particularity measure for functional characterization of gene sets using gene ontology We propose a new approach to compute gene set particularities based on the information conveyed by Gene Ontology terms. A GO term informativeness can be computed using either its information content based on the term frequency in a corpus, or a function of the term's distance to the root. We demonstrated that the combination of semantic similarity and semantic particularity measures was able to identify genes with particular functions from among similar genes. This differentiation was not recognized using only a semantic similarity measure. [*C. Bettembourg, O. Dameron*] [17] [Online publication]

Integrating GALAXY workflows in a metadata management environment New tools are needed to enable the quick design and the intensive parallel execution of bioinformatics processes. Therefore, we proposed a new dataflow-oriented workflow management system dedicated to intensive bioinformatics tasks. We worked on the interoperatibility of bioinformatics workflows using a model-driven approach. Our results enable new import / export capabilities beetwen multiple workflow management environnements and insights to create a unique shared workflow model. [*O. Dameron, F. Moreews (Genscale team), Y. Le Bras (GenOuest platform), C. Monjeaud (GenOuest platform), O. Collin (GenOuest platform)*][36]

6.3. Time-series and asymptotic dynamics

Participants: Anne Siegel, Jérémie Bourdon, Jeanne Cambefort, Damien Eveillard, Vincent Picard, Nathalie Théret, Santiago Videla.

Reasoning on the response of logical signaling networks with boolean models A series of papers and a PhD thesis focused on modeling the response of logical signaling networks by means of automated reasoning using ASP. In this context, a crucial issue is automatic learning of logical networks from partial observations of input/output behaviours, in order to achieve unbiased and robust discoveries. Experiments showed that many networks can be compatible with a given set of experimental observations. In a review chapter, we first discuss how ASP can be used to exhaustively enumerate all these logical networks. Next, in order to gain control over the system, we look for intervention strategies that force a set of target species into a desired steady state.

Finally, we discuss the usage of ASP for solving the aforementioned problems and the novelty of our approach with respect to existing methods. [*S. Videla, A. Siegel, J. Nicolas*] [23], [38], [12] [Online publication]

Integrative modeling framework for signaling networks based on guarded transitions models We develop a new non-ambiguous formal interpretation of signaling pathways as discrete dynamic models. The resulting language, Computer-Aided Design for BIOlogical Models (CADBIOM), is based on a simplified version of guarded transitions in which we introduced temporal parameters for each transition to manage competition and cooperation between parts of the models . Tools for simulation and model checking analyses using the formal Cadbiom language have been developed (http://cadbiom.genouest.org). Using CADBIOM, we built the first discrete model of TGF-β signaling networks by automatically integrating the 137 human signaling maps from the Pathway Interaction Database into a single unified dynamic model. Temporal property-checking analyses of 15934 trajectories that regulate 145 TGF-$beta$ target genes reveal the association of specific pathways with distinct biological processes. [*G Andrieux, M Le Borgne, N. Théret*] [15] [Online publication]

Exploring metabolism flexibility in complex organisms 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. [*O. Abdou-Arbi, J. Bourdon, A. Siegel*] [14] [Online publication]

Multivariate Normal Approximation for the Stochastic Simulation Algorithm: limit theorem and applications We prove a central limit theorem on the asymptotic stochastic dynamical behavior of the outputs of a reaction network under certain steady-state assumptions. We present multiple applications including a constraints-based approach to verify asymptotic properties on the output moments without prior knowledge about the kinetic parameters. [*V. Picard, A. Siegel, J. Bourdon*] [33] [Online publication]

Numeric model for initiation of translation in sea-urchin We use a numeric-based modeling approach to study the regulation of protein synthesis following fertilization in sea urchin. This approach based on parcimonious modelling evidenced that two processes are required to explain experimental data: a destabilization of eIF4E:4E-BP complex and a great stimulation of the 4E-BP-degradation mechanism, both rapamycin-sensitive [*A. Siegel, J. Bourdon*] [21] [Online publication]

6.4. Sequence annotation

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

Automated Enzyme Classification by Formal Concept Analysis Guessing enzyme's functional activity from its sequence is a crucial task that can be approached by comparing the new sequences with those of already known enzymes labeled by a family class. This task is difficult because the activity is based on a combination of small sequence patterns and sequences greatly evolved over time. We have designed a classifier based on the identification of common subsequence blocks between known and new enzymes and the search of formal concepts built on the cross product of blocks and sequences for each class. Since new enzyme families may emerge, it is important to propose simultaneously a first classification of enzymes that cannot be assigned to a known family. Formal Concept Analysis offers a nice framework to set this task as an optimization problem on the set of concepts. The classifier has been tested with success on a particular set of enzymes present in a large variety of species, the haloacid dehalogenase (HAD) superfamily. [*F. Coste, G. Garet, J. Nicolas*] [28], [10]

A bottom-up efficient algorithm learning substitutable languages from positive examples Based on Harris's substitutability criterion, the recent definitions of classes of substitutable languages have led to interesting polynomial learnability results for expressive formal languages. These classes are also promising for practical applications: in natural language analysis, because definitions have strong linguisitic support, but also in biology for modeling protein families, as suggested in our previous study introducing the class of

local substitutable languages. But turning recent theoretical advances into practice badly needs truly operable algorithms. We present here an efficient learning algorithm, motivated by intelligibility and parsing efficiency of the result, which directly reduces the positive sample into a small non redundant canonical grammar of the target substitutable language. Thanks to this new algorithm, we have been able to extend our experimentation to a complete protein dataset confirming that it is possible to learn grammars on proteins with high specificity and good sensitivity by a generalization based on local substitutability. [*F. Coste, G. Garet, J. Nicolas*] [29], [10]

Logol: Expressive Pattern Matching in sequences. Application to Ribosomal Frameshift Modeling Logol consists in both a language for describing biological patterns, and an associated parser for effective pattern search in sequences (RNA, DNA or protein). The Logol language, based on an high level grammatical formalism (String Variable Grammars), allows to express flexible patterns (with mispairings and indels) composed of both sequential elements (such as motifs) and structural elements (such as repeats or pseudoknots). Its expressive power allows the design of sophisticated patterns such as the signature of "-1 programmed ribosomal frameshifting" (PRF) events in messenger RNA sequences. A PRF signature is a complex model composed of a slippery site followed by a pseudoknot located in a specific part of the sequence, which provides a good illustration of the Logol language power. [*C. Belleannée, J. Nicolas, O. Sallou (GenOuest platform)*] [27] [Online publication]

Identifying distant homologous viral sequences in metagenomes using protein structure information It is estimated that marine viruses daily kill about 20% of the ocean biomass. Identifying them in water samples is thus a biological issue of great importance. The metagenomic approach for virus identification is a challenging task since their sequences carry a lot of mutations and are very difficuly to identify by standard homology searches. The PEPS VAG project aims at establishing a novel methodology that uses structures of proteins as extra-information in order to annotate metagenomes without relying on sequence homology. In the context of the first experiments made on the metagenome of station 23 of the TARA Ocean Project, we used the structures of capsid proteins to infer the sequence signature of their fold, in order to find them in the metagenome. This work presents the methodology, the first experiments and the on-going improvements. [*C. Galiez, F. Coste*] [35]

Computational Protein Design: trying an Answer Set Programming approach to solve the problem The problem of *Computational Protein Design* aims at finding the best protein conformation to perform a given task. This problem can be reduced to an optimization problem, looking for the minimum of an energy function depending on the amino-acid interactions in the protein. The CPD problem may be easily modeled as an ASP program but a practical implementation able to work on real-sized instances has never been published. We have raised the main source of difficulty for current ASP solvers and ran a series of benchmarks highlighting the importance of finding a good upper bound estimation of the target minimum energy to reduce the amount of combinatorial search. Our solution clearly outperforms a direct ASP implementation without this estimation and has comparable performances with respect to SAT-based approaches. It remains less efficient than a recent approach by cost function networks, showing there still exists some place for improving the optimization component in ASP with more dynamical strategies. [*J. Nicolas, H. Bazille*] [34]

Searching for Optimal Orders for Discretized Distance Geometry The Molecular Distance Geometry Problem (MDGP) is the problem of finding the possible conformations of a molecule by exploiting available information on some distances between pairs of its atoms. When some assumptions are satisfied, the MDGP can be discretized, so that the search domain of the problem becomes a tree where each node corresponds to a candidate position for an atom. The search tree can be efficiently explored by using an *interval* Branch & Prune (iBP) algorithm that can potentially enumerate all feasible conformations. In this context, the order given to the atoms of the molecule plays an important role, because it allows the discretization assumptions to be satisfied, and it also impacts the computational cost of the iBP algorithm. We have proposed a new discretization order for protein backbones based on the optimization of certain criteria for a faster exploration of the discretized search domain. To this aim, we express the search for optimal orders by a set of logical constraints in ASP. Our comparison with previously proposed orders for protein backbones shows that this new discretization order makes iBP perform better. [*J. Nicolas, A. Muccherino (Genscale Team)*] [43]

From analogical proportions in lattices to proportional analogies in formal concepts We provided an attempt at bridging formal concept analysis and the modeling of analogical proportions (i.e., statements of the form "a is to b as c is to d"). A suitable definition for analogical proportions in non distributive lattices is proposed and then applied to concept lattices. This enables us to compute what we call proportional analogies. In addition, we define the locally maximal subwords and locally minimal superwords common to a finite set of words. We also define the corresponding sets of alignments. We show that the constructed family of sets of alignments has the lattice structure. The study of analogical proportion in lattices gives hints to use this structure as a machine learning basis, aiming at inducing a generalization of the set of words. [*L. Miclet*] [32], [37]

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, Vincent Picard, Santiago Videla.

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

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). In addition, the team participates to a collaboration program with the Biocore and Ange teams, together with Ifremer-Nantes, focused on the understanding on micro-algae (thesis of Julie Laniau).

7.1.3. Regional partnership in agriculture and bio-medical domains

Participants: Aymeric Antoine-Lorquin, Catherine Belleannée, Charles Bettembourg, François Coste, Jean Coquet, Olivier Dameron, Victorien Delannée, Jacques Nicolas, Anne Siegel, Valentin Wucher, Nathalie Théret.

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 V. Wucher [13] was supported by collaborations with the IGEPP laboratory. The post-doc of Charles Bettembourg nows stregthens these collaborations. This collaboration is also reinforced by collaboration within ANR contracts (MirNadapt, FatInteger).

We also have a strong and long term collaboration in the bio-medical domain, namely with the IRSET laboratory at Univ. Rennes 1/Irset, 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.

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, Marie Chevallier, Damien Eveillard, Gaëlle Garet, Jacques Nicolas, 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. 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.

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, Yann Guitton, 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, Olivier Dameron, Vincent Picard, Sylvain Prigent, Nathalie Théret, 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 Nov. 2014. Teams involved include LINA (Nantes), I3S (Nice), DIMPP (Montpellier), Contraintes/Lifeware 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, 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 Mirnadapt

Participants: Jacques Nicolas, Catherine Belleannée, Anne Siegel, Olivier Dameron, Valentin Wucher, Charles Bettembourg.

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 was strengthened by a co-tutored PhD thesis (V. Wucher) defended in Nov. 2014 [13]. 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.

7.2.3.3. ANR Samosa

Participants: Jacques Nicolas, Catherine Belleannée, Anne Siegel, Aymeric Antoine-Lorquin, Jérémie Bourdon, François Coste.

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.

7.2.4. Programs funded by research institutions

7.2.4.1. ADT Complex-biomarkers

Participants: Jeanne Cambefort, Guillaume Collet, Marie Chevallier, 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).

7.2.4.2. 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.

7.2.4.3. 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 - 2016

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 2016. 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 (2012) and the co-supervised 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 collaboration within the framework of the CIRIC center. In 2014, Marie Chevallier joined the team as an engineer to reinforce software resulting from common collaborations.

7.4.3. Participation in other International Programs

7.4.3.1. International joint supervision of PhD

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

- **Algeria**. Badji Mokhtar - Annaba University [M. Zekri]
- **Austria**. Graz university [M. Weltzer]
- **Chile**. Centro de Modelimiento Matematico, Santiago [A. Maass, P. Bordron, M.P. Cortez]
- **Germany**. Department of Computer Science, Potsdam [T. Schaub]
- **Germany**. Frei Universitat Berlin [A. Bockmayr]

7.5.1.1. Internships

Francisco Dorr

Date: Mar 2014 - Aug 2014

Institution: Universidad de Buenos Aires (Argentina)

7.5.2. Visits to International Teams

7.5.2.1. Shorts visits

- **Chile**. Centro de Modelimiento Matematico, Santiago. *Applications of ASP*. Nov. 2014 (1 to 2 weeks) [J. Bourdon, M. Chevallier, D. Eveillard, A. Siegel]

7.5.2.2. Explorer programme

Prigent Sylvain

Date: Mar 2014 - Apr 2014

Institution: FUB (Germany)

Videla Santiago

Date: Mar 2014 - May 2014

Institution: University of Potsdam (Germany)

Picard Vincent

Date: Sep 2014 - Nov 2014

Institution: The University of Tokyo, Japanese-French Laboratory for Informatics(Japan)

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific events selection

8.1.1.1. Members of conference program committee

- ICGI International Conference on Grammatical Inference [F. Coste, Reviewer and member of Steering Committee].
- PRIB International Conference on Pattern Recognition in Bioinformatics [A. Siegel].
- Program committee of Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2014) [O. Dameron].

8.1.2. Journal

8.1.2.1. Member of editorial board

- Academic editor: Plos One [J. Bourdon]

8.1.2.2. Reviewer

- F. Coste: Theoretical Computer Science.
- O. Dameron: Bioinformatics, Cancer informatics, Journal of Biomedical Informatics, Journal of Biomedical Semantics.
- A. Siegel : Nature, Bulletin of the Belgium Math. Society, BMC Systems Biology.

8.1.3. Participation to council and advisory boards

- Scientific Advisory Board of GDR BIM " Molecular Bioinformatics"[J. Nicolas].
- Member of SCAS (Service Commun d'Action Sociale) of Univ. Rennes 1 [C. Belleannée].
- Member of the IRISA laboratory council [F. Coste].
- Member of the Inria Rennes center council [A. Siegel].
- Member of the the Operational Legal and Ethical Risk Assessment Committee (COERLE) at Inria.
- Scientific Advisory Board of Biogenouest [J. Bourdon, J. Nicolas, A. Siegel].
- Expertise for ANR call [A. Siegel]
- Expertise for national call "Biologie des systèmes appliqués au Cancer" [A. Siegel]
- Expertise for ARP "MERMED : Adaptation aux changements environnementaux en mer Méditerranée : quelles recherches et quels partenariats ?" [website] [A. Siegel]
- Recruitment committees: assistant professor (LRI, Orsay) [A. Siegel, F. Coste], assistant professor (Univ. Rennes) [F. Coste], professor (Univ. Bordeaux) [A. Siegel].

8.1.4. Organization of local meetings

- **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].
- Organization of Irisa department "Data and knowledge management"'s seminar [F. Coste].
- **Workshop. Théorie des réseaux booléens et ses applications en biologie** A one-week workshop on boolean networks was organized in Nice in Nov. 2014. It gathered 20 researchers in bioinformatics, computer science and mathematics about several aspects of boolean networks properties [website]

8.2. Teaching - Supervision - Juries

8.2.1. Course and track responsibilities

F. Coste is coordinator of the track "From Data to Knowledge: Machine Learning, Modeling and Indexing Multimedia Contents and Symbolic Data" of the Master by research in Computer Science (2nd year), University of Rennes 1, France.

F. Coste is coordinator of the course “Extracting knowledge from symbolic data sequences” of the Master by research in Computer Science (2nd year), University of Rennes 1, France.

8.2.2. Teaching

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

Licence: C. Belleannée, bureautique et C2i, 40h, L1 informatique, Rennes1, France.

Licence: G. Garet, Algorithmique, 22h, L3 informatique, Rennes1 France.

Licence: G. Garet, Analyse de données, 12h, L3 Miage, Rennes1 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, Probability theory, 24h, L3, ENS Rennes, France

Licence: C. Galiez, Bureautique, 16h, L1 informatique, Rennes1 France

Master: C. Bettembourg, principes de programmation et d’algorithmique, 18h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

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

Master: G. Collet, Programmation Python, 20h, M1 BioInformatique, Univ. Rennes 1, France

Master: O. Dameron, gestion de projets en informatique, 29.5h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, principes de programmation et d’algorithmique, 12.5h, 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, techniques de recherche documentaire, 2h, M1 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, Bases de mathématiques, probabilités et statistiques, 3h, M1 santé publique, Univ. Rennes 1, France.

Master: C. Galiez, Compilation, 48h, M1 informatique, Rennes1 France

Master: V. Picard, Formal methods for safe development, 16h, M1, Univ. Rennes 1, France

Master: V. Picard, Agrégation de mathématiques option D, 15h, M1, ENS Rennes/Univ. Rennes 1, 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, modélisation des connaissances et bio-ontologies, 36h, M2 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, organisation des oraux de stages, 8h, M2 bioinformatique et génomique, Univ. Rennes 1, France.

Master: O. Dameron, E-santé et réseaux hospitaliers, 8h, ESIR3, Univ. Rennes 1, France.

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

Doctorat: A. Siegel, Tilings and Symbolic Dynamics, 8h, School "Representing streams II" at Lorentz center, Leiden, Netherlands.

Doctorat: J. Nicolas, Answer Set programming, 6h, Ecole Doctorale Matisse, Univ. Rennes 1, France

8.2.3. Supervision

PhD : Gaëlle Garet, *Discovery of enzymatic functions in the framework of formal languages*, 16 Dec. 2014,, supervised by J. Nicolas and F. Coste. [10].

PhD : Sylvain Prigent, *Complétion combinatoire pour la reconstruction de réseaux métaboliques, et application au modèle des algues brunes Ectocarpus siliculosus*, 14 Nov. 2014, supervised by A. Siegel and T. Tonon (UMR 7150, station biologique de Roscoff) [11]

PhD : Santiago Videla, *Reasoning on the response of logical signaling networks with answer set programming*, 7 Jul. 2014, supervised by A. Siegel and T. Schaub (Potsdam univ) [12].

PhD : Valentin Wucher, *Modeling of a gene network between mRNAs and miRNAs to predict gene functions involved in phenotypic plasticity in the pea aphid*, 3 Nov. 2014, supervised by J. Nicolas and D. Tagu (Inra).[13]

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 : Clovis Galiez, *Syntactic modelling of protein structure.*, started in Oct. 2012, supervised by F. Coste and J. Nicolas.

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

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.

8.2.4. Juries

- *Member of Ph-D thesis jury*. M. Folschette, Ecole Centrale Nantes [A. Siegel, rapporteure]. B. Le Gloannec, Univ. Orléans [A. Siegel, présidente]. J. Scicluna, Université de Nantes [F. Coste].

8.2.5. Seminars

- V. Picard. *Multivariate Normal Approximation for the Stochastic Simulation Algorithm: limit theorem and applications*. I3S MDSC Seminar. Sophia Antipolis, France (May. 2014)
- A. Siegel. *modelling and integrating heterogeneous information about the response of a biological system with ASP*. LRI, Orsay, France (Jun. 2014)
- A. Siegel. *AuReMe Integrative method for Automatic Reconstruction of Metabolic network*. LINA, Nante, France (Jul. 2014)
- A. Siegel. *Modéliser et intégrer des informations hétérogènes sur la réponse d'un système biologique*. Séminaire interdisciplinaire MEB, Marseille, France (Oct. 2014)
- G. Collet. *Genome Assembly on a Raspberry Pi*. fOSSa 2014 (Nov. 2014)
- A. Siegel. *Méthodes de programmation logique (ASP) pour apprendre et contrôler la réponse de réseaux de signalisation*. Workshop "Théorie des réseaux booléens et ses applications en biologie", Nice, France (Nov. 2014)
- A. Siegel. *Reconstruire un réseau métabolique à partir de différentes sources d'informations et données*. Journée scientifique BioGenOuest - Axe Analyse structurale et métabolomique, Nantes, France (Dec. 2014)
- A. Siegel. *Modéliser et intégrer des informations hétérogènes à large-échelle sur la réponse d'un système biologique*. Journées biologie intégrative, Lille, France (Dec. 2014)

- O. Dameron *OWL model of eligibility criteria compatible with partially-known information*. Institut des Systèmes Complexes, Paris, France (Dec. 2014)
- V. Picard. *Multivariate Normal Approximation for the Stochastic Simulation Algorithm: limit theorem and applications*. LIFL Biocomputing Seminar. Lille Univ, France (Dec. 2014)
- O. Dameron *Knowledge-based selection of candidate metabolic networks*. Séminaire Biosticker, LINA Nantes France (Dec. 2014)

8.2.5.1. Internships

- Internship, from April until June 2014. Supervised by G. Collet. Student: Efflam Lemaillet. Subject: Méthode de comparaison de protéines sans alignement.
- Internship, from April until June 2014. Supervised by F. Coste. Student: Francisco Dorr. Subject: Compressing (genomic) sequences by grammar inference.
- Internship, from April until June 2014. Supervised by O. Dameron and N. Theret. Student: DominiqueMias-Lucquin. Subject: Analysis of TGF-β signalling network trajectories
- Internship, from April until June 2014. Supervised by O. Dameron and G. Collet. Student: LoïcBourgeois. Subject: Metabolic pathway representation in RDF : converting Reactome into BioPAX
- Internship, from Feb. until June 2014. Supervised by J. Nicolas. Student: Hugo Bazille. Subject: Protein design: a NP-hard problem in bioinformatics
- Internship, from March until Sep. 2014. Supervised by C. Belleannée. Student: Malo Le Boulch. Subject:
- Internship, from March until June 2014. Supervised by C. Belleannée. Student: Lea Flechon. Subject: Identification in silico du site de fixation à l'ARNm de la protéine PTBP1 : découverte de motifs par utilisation d'un outil existant via des données CLIP-seq publiques
- Internship, from April until June 2014. Supervised by V. Picard and A. Siegel. Student: Thibault Etienne. Subject: Simulation stochastique quantitative pour la modélisation de systèmes écologiques
- Internship, from January until June 2014. Supervised by A. Siegel and N. Theret. Student: Victorien Delannée. Subject: Modélisation du métabolismes des xénobiotiques de type amines hétérocycliques aromatique (AHA)
- Internship, October 2014. Supervised by F. Coste. Ph D. student: Meriem Zekri. Subject: Modelling GPCR proteins with Protomata-Learner

8.3. Popularization

- *Revue Math/info Tangente* Nous avons participé à un numéro spécial de cette revue destinée aux lycées avec un article de bioinformatique sur les langages des molécules du vivant illustrant l'intérêt de la modélisation par langages formels dans ce cadre. [*J. Nicolas, C. Belleannée, F. Coste*] [41]
- *Rencontre Inria Industrie* "Bio-informatique et outils numériques pour les produits de santé" Lyon (Feb. 2014) [*F. Coste, A. Antoine-Lorquin, A. Siegel, S. Videla*]
- *FrenchTech Rennes* Popularization of genome assembly on a Raspberry Pi at a FrenchTech event in Rennes (Dec. 2014) [*G. Collet*].
- *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, network visualization. [*G. Collet, S. Prigent, O. Dameron*]. [more info].
- *Fête de la science (LINA, Nantes)* During the 17th of october, 250 students discovered bioinformatics, genome assembly, metabolomics and algorithmic by practicing games and tutorials made by our team with the help of Julien Gras, Erwan Delage and Stephanie Noguet (LINA) [*J. Bourdon, D. Eveillard, Y. Guitton*].

- *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].
- *GNU/Linux Magazine* Article "La bioinformatique avec Biopython", Hors-série 73, pp 84–97. 2014. [*O. Dameron*] [40]

9. Bibliography

Major publications by the team in recent years

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] G. GARET. *Classification and characterization of enzymatic families with formal methods*, Université de Rennes 1, December 2014, https://hal.inria.fr/tel-01096916

[11] S. PRIGENT. *Combinatorial completion for the reconstruction of metabolic networks, and application to the brown alga model Ectocarpus siliculosus*, Université de Rennes 1, November 2014, https://hal.inria.fr/tel-01093287

[12] S. VIDELA. *Reasoning on the response of logical signaling networks with answer set programming*, Université Rennes 1 ; Universität Postdam (Allemagne), July 2014, https://tel.archives-ouvertes.fr/tel-01070436

[13] 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

Articles in International Peer-Reviewed Journals

[14] O. ABDOU-ARBI, S. LEMOSQUET, V. 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

[15] G. ANDRIEUX, M. LE BORGNE, N. THÉRET. *An integrative modeling framework reveals plasticity of TGF-β 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

[16] V. BERTHÉ, T. JOLIVET, A. SIEGEL. *Connectedness of fractals associated with Arnoux–Rauzy substitutions*, in "RAIRO: Informatique Théorique et Applications / RAIRO: Theoretical Informatics and Applications", 2014, vol. 48, n^{o} 3, pp. 249-266 [*DOI :* 10.1051/ITA/2014008], https://hal.inria.fr/hal-01079727

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

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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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)", June 2014 [*DOI :* 10.1016/J.TCS.2014.06.022], https://hal.inria.fr/hal-01058610

Invited Conferences

[24] A. SIEGEL, C. GUZIOLOWSKI, S. VIDELA, T. SCHAUB. *Improving robustness in the study of logical model of signalling networks with answer set programming*, in "ECCB 2014 - Workshop on Logical Modelling and Analysis of Cellular Networks", Strasbourg, France, September 2014, https://hal.inria.fr/hal-01095607

[25] A. SIEGEL. *Using ASP to integrate large-scale heterogeneous information about the response of a biological system* , in "5th Workshop on Logic and Systems Biology (associated with CSL/LICS 2014)", VIenna, Austria, July 2014, https://hal.inria.fr/hal-01095609

International Conferences with Proceedings

[26] 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

[27] C. BELLEANNÉE, O. SALLOU, J. NICOLAS. *Logol: Expressive Pattern Matching in sequences. Application to Ribosomal Frameshift Modeling*, in "PRIB - 9th IAPR International Conference on Pattern Recognition in Bioinformatics", 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

[28] 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

[29] F. COSTE, G. GARET, J. NICOLAS. *A bottom-up efficient algorithm learning substitutable languages from positive examples*, in "ICGI (International Conference on Grammatical Inference)", Kyoto, Japan, A. CLARK, M. KANAZAWA, R. YOSHINAKA (editors), The 12th International Conference on Grammatical Inference, September 2014, vol. 34, pp. 49-63, https://hal.inria.fr/hal-01080249

[30] 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

[31] N. MAILLET, G. COLLET, T. VANNIER, D. LAVENIER, P. PETERLONGO. *COMMET: comparing and combining multiple metagenomic datasets*, in "IEEE BIBM 2014", Belfast, United Kingdom, November 2014, https://hal.inria.fr/hal-01080050

[32] L. MICLET, N. BARBOT, H. PRADE. *From analogical proportions in lattices to proportional analogies in formal concepts*, in "ECAI - 21th European Conference on Artificial Intelligence", Prague, Czech Republic, August 2014, https://hal.inria.fr/hal-01000314

[33] V. PICARD, A. SIEGEL, J. BOURDON. *Multivariate Normal Approximation for the Stochastic Simulation Algorithm: limit theorem and applications*, in "SASB - 5th International Workshop on Static Analysis and Systems Biology", Munchen, Germany, 2014, https://hal.inria.fr/hal-01079768

Conferences without Proceedings

[34] H. BAZILLE, J. NICOLAS. *Computational Protein Design: trying an Answer Set Programming approach to solve the problem*, in "10th Workshop on Constraint-Based Methods for Bioinformatics (WCB'14)", Lyon, France, Nicos Angelopoulos (Imperial College, UK) , Simon de Givry (MIAT-INRA, France), September 2014, https://hal.inria.fr/hal-01063030

[35] M. BOCCARA, M. CARPENTIER, J. CHOMILIER, F. COSTE, C. GALIEZ, J. POTHIER, A. VELUCHAMY. *Identifying distant homologous viral sequences in metagenomes using protein structure information*, in "ECCB'14 Workshop on Recent Computational Advances in Metagenomics", Strasbourg, France, September 2014, https://hal.archives-ouvertes.fr/hal-01090987

[36] F. MOREEWS, Y. LE BRAS, O. DAMERON, C. MONJEAUD, O. COLLIN. *Integrating GALAXY workflows in a metadata management environment*, in "Galaxy Community Conference", Baltimore, United States, July 2014, https://hal.inria.fr/hal-01093058

Scientific Books (or Scientific Book chapters)

[37] L. MICLET, N. BARBOT, B. JEUDY. *Analogical Proportions in a Lattice of Sets of Alignments Built on the Common Subwords in a Finite Language*, in "Computational Approaches to Analogical Reasoning: Current Trends, Studies in Computational Intelligence", H. PRADE, G. RICHARD (editors), Springer-Verlag Berlin Heidelberg, 2014, pp. 245-260 [*DOI :* 10.1007/978-3-642-54516-0_10], https://hal.inria.fr/hal-00974656

[38] 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

[39] 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

Scientific Popularization

[40] O. DAMERON, G. FARRANT. *La bioinformatique avec Biopython*, July 2014, 13 p. , GNU/Linux Magazine, hors-série 73, https://hal.inria.fr/hal-01095475

[41] J. NICOLAS, C. BELLEANNÉE, F. COSTE. *Le langage des molécules du vivant*, in "Bibliothèque Tangente", 2014, n° 52, 8 p. , https://hal.inria.fr/hal-01100051

Other Publications

[42] H. BAZILLE. *Protein design: a NP-hard problem in bioinformatics*, IRISA-Inria, Campus de Beaulieu, 35042 Rennes cedex, 2014, 42 p. , http://dumas.ccsd.cnrs.fr/dumas-01088787

[43] D. GONÇALVES, J. NICOLAS, A. MUCHERINO. *Searching for Optimal Orders for Discretized Distance Geometry*, October 2014, Proceedings of Many Faces of Distances (MDF14), https://hal.inria.fr/hal-01093072

[44] Y. LE BRAS, A. ROULT, C. MONJEAUD, B. MATHIEU, O. QUENEZ, H. CLAUDIA, O. SALLOU, A. BRETAUDEAU, O. COLLIN. *e-Science in France, a Life science Western story*, June 2014, Galaxy Community Conference (GCC2014), https://hal.inria.fr/hal-01102432

References in notes

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

[46] 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° 1, 146 p. [*DOI :* 10.1186/1752-0509-4-146], http://www.hal.inserm.fr/inserm-00984711

[47] 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

[48] 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

[49] 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° 2, pp. 93-115

[50] 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

[51] 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° 1-2, pp. 7–11

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

[53] 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

[54] 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

[55] 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

[56] 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

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

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

[59] 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

[60] 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