IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Activity Report 2012

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 and Bioinformatics

Table of contents

Team DYLISS

Keywords: Computational Biology, Genetic Networks, Network Dynamics, Reasoning, Machine Learning, Markovian Model

The Dyliss project was created as a spin-off of the former Symbiose project. It is a team from Inria-Rennes, Université de Rennes 1 and CNRS.

Creation of the Team: January 01, 2012 .

1. Members

Research Scientists

Anne Siegel [Team leader, Senior Researcher Cnrs, HdR]
François Coste [Junior researcher, Inria]
Jacques Nicolas [Senior researcher, Inria, HdR]

Faculty Members

Catherine Belleannée [Associate Professor, Univ. Rennes 1]
Michel Le Borgne [Associate Professor, Univ. Rennes 1]
Laurent Miclet [Professor, Emeritus, Univ. Rennes 1]

External Collaborators

Jérémie Bourdon [Associate Professor, Univ. Nantes, HdR]
Damien Eveillard [Associate Professor, Univ. Nantes]

Engineers

Guillaume Collet [non permanent contract, ANR Idealg, since oct. 2012]
Claudia Hériveau [non permanent junior engineer, ADT Inria]

PhD Students

Andres Aravena [Co-supervision. Main institution: University of Chile]
Oumarou Abdou-Arbi [MENRT]
Geoffroy Andrieux [MENRT. Cosupervision. Main institution: IRSET/Rennes 1]
Clovis Galiez [Inria Cordi-S, since oct. 2012]
Gaëlle Garet [Région/Inria]
Vincent Picard [ENS Rennes, since sept. 2012]
Sylvain Prigent [MENRT]
Santiago Videla [CNRS/ANR Biotempo]
Valentin Wucher [Région Bretagne/INRA. Cosupervision. Main institution: IGEPP/INRA]

Post-Doctoral Fellows

Pierre Blavy [INRA, ASC]
Sven Thiele [Inria, since apr. 2012]

Administrative Assistant

Marie-Noëlle Georgeault [Assistant, Inria]

2. Overall Objectives

2.1. Highlights of the Year

- François Coste was the co-chair of the French conference in bioinformatics (JOBIM) which was organized in Rennes in July 2012.
- Matthias Gallé, a former PhD in the team, won the accessit thesis prize from AFIA. This work followed by F. Coste has been achieved in the framework of a cooperation with Universidad Nacional de Cordoba, thanks to a MinCYT-Inria program [14].
- Santiago Videla won a best paper award at the conference CMSB [19] [1]. This work implies a cooperation with EBI (UK) together with universities of Heidelberg, Potsdam and Padova.

BEST PAPER AWARD :
[19] **Revisiting the Training of Logic Models of Protein Signaling Networks with a Formal Approach based on Answer Set Programming in CMSB - 10th Computational Methods in Systems Biology 2012**. S. VIDELA, C. GUZIOLOWSKI, F. EDUATI, S. THIELE, N. GRABE, J. SAEZ-RODRIGUEZ, A. SIEGEL.

2.2. 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 response of living species capable of facing extreme environments. 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 (knowledge representation, constraints programming) to take into account information on physiological responses of the studied species under various constraints (knowledge representation, constraint programming, multi-scale analysis of dynamical systems) as well as genetic information from their long-distant cousins (machine learning).

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, which needs to better formalize the experimentation loop.

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 mainly observed at the physiological 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 families of genomic products implied in the living system response. 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.

[1] http://sites.brunel.ac.uk/cmsb2012

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. Scientific Foundations

3.1. Knowledge representation with constraint programming

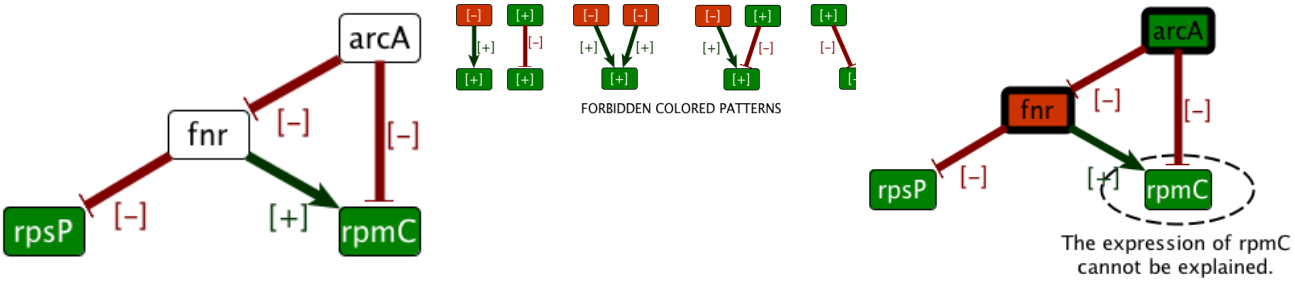
Biological networks are built with data-driven approaches aiming at translating genomic information into a functional map. Most methods are based on a probabilistic framework which defines a probability distribution over the set of models. The reconstructed network is then defined as the most likely 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. Common properties to all solutions are considered as a robust information about the system, as they are independent from the choice of a single solution to the optimization problem[6].

Solving these computational issues requires addressing NP-hard qualitative (non-temporal) issues, based on a notion of causality. We have developed a long-term collaboration with Potsdam University in order to use a logical paradigm named **Answer Set Programming** [27], [30] to solve these optimization issues. Applied on transcriptomic or cancer networks, our methods identified which regions of a large-scale network shall be corrected [1], and proposed robust corrections [5]. 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 optimization issues explored in systems biology.

Using these technologies requires to revisit and reformulate optimization problems at hand in order both to decrease the search space size in the grounding part of the process and to optimize 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



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.

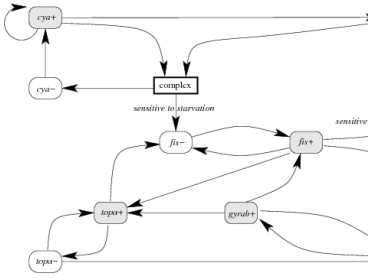
Step 3. The model allows both the prediction of values (e.g. for fnr in the figure) and the detection of contradictions (e.g. the expression level of rpmC is inconsistant with the regulation in the graph).

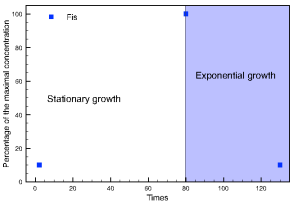
Step 4. Excerpt from the ASP program.

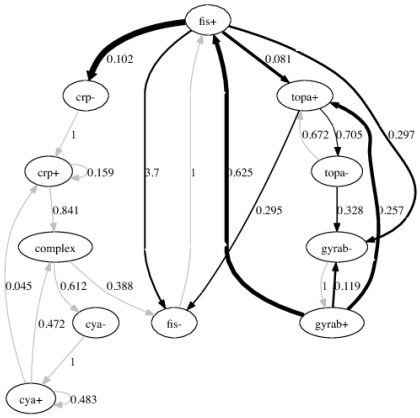
```
vertex(fnr).                    1{labelV(I ,+;-)}1 :- vertex(I).        receive(I,+) :- labelE(J,I,S), labelV(J,S).
edge(fnr,rpsP).                 labelV(I ,S) :- observedV(I,S).         receive(I,-) :- labelE(J,I,S), labelV(J,T), S≠T.
observedE(fnr,rpsP,-).          1{labelE(J,I,+;-)}1 :- edge(J,I).       :- labelV (I,S), not receive(I,S).
observedV(rpsP,-).              labelE(J,I,S) :- observedE(J,I,S).
```

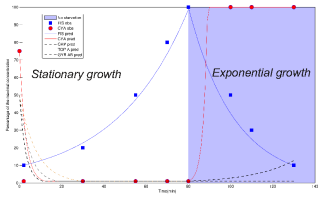
Figure 1. Excerpt from the ASP program identifying which expression of non-observed nodes (white nodes) are fixed by partial observations and rules derived from the system dynamics. The logical approach is flexible enough to model in a single framework network characteristics (products, interactions, partial information on signs of regulations and observations) and static rules about the effects of the dynamics of the system. Extensions of this framework include the exhaustive search for system repair or more constrained dynamical rules [6], [5] .

mechanistic and probabilistic modeling [2]. The system is modeled by an Event Transition Graph, that is, a **Markovian qualitative description of its dynamics** together with quantitative laws which describe the effect of the dynamic transitions over higher scale quantitative measurements. Then, a few time-series quantitative 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 [29]. This property can be derived into constraints to describe the set of admissible weighted Markov chains whose asymptotic behavior agrees with the quantitative measures at hand. A precise study of this constrained space via local search optimization emphasizes the most important discrete events that must occur to reproduce the information at hand. These methods have been validated on the *E. coli* regulatory network benchmark. We now plan to apply these techniques to reduced networks representing the main pathways and actors automatically generated from the integrative methods developed in Axis 1. 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.









Input data. Qualitative description of the system dynamics at the transcription level (interaction graph) and 3 concentration measurements of the fis protein (population scale).

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.

Prediction of the Cya protein concentration (red curve) fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in E. Coli response to nutritional stress.

Figure 2. Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems. Identification of key interactions [2].

3.3. Grammatical inference and highly expressive structures

Our main field of expertise in **machine learning** concerns grammatical models with a long -term know-how in finite state automata learning. By introducing a similar fragment merging heuristic approach, we have proposed an algorithm that learns successfully automata modeling families of (non homologous) functional families of proteins [4], leading to a tool named Protomata-learner. As an example, this tools allows to properly model the multi-domain function of the protein family TNF, which is impossible with other existing probabilistic-based approach (see Fig. 3). Our future goal is to demonstrate the relevance of formal language theory by addressing the question of enzyme prediction, from their genomic or protein sequences, aiming at better sensitivity and specificity. As enzyme-substrate interactions are very specific central relations for integrated genome/metabolome studies and are characterized by faint signatures, we shall rely on models for active sites involved in cellular regulation or catalysis mechanisms. This requires to build models gathering both structural and sequence information is order to describe (potentially nested or crossing) long-term dependencies such as contacts of amino-acids that are far in the sequence but close in the 3D protein folding. We wish to extend our expertise towards inferring Context-Free Grammars including the topological information coming from the structural characterization of active sites.

Moving forward to context-free grammars instead of regular patterns increases **parsing** complexity. 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 [31]. An extended formalism, String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [34], [33]. 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 [9]. 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.

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.

4.2. Biological data integration

Axis 1 (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).

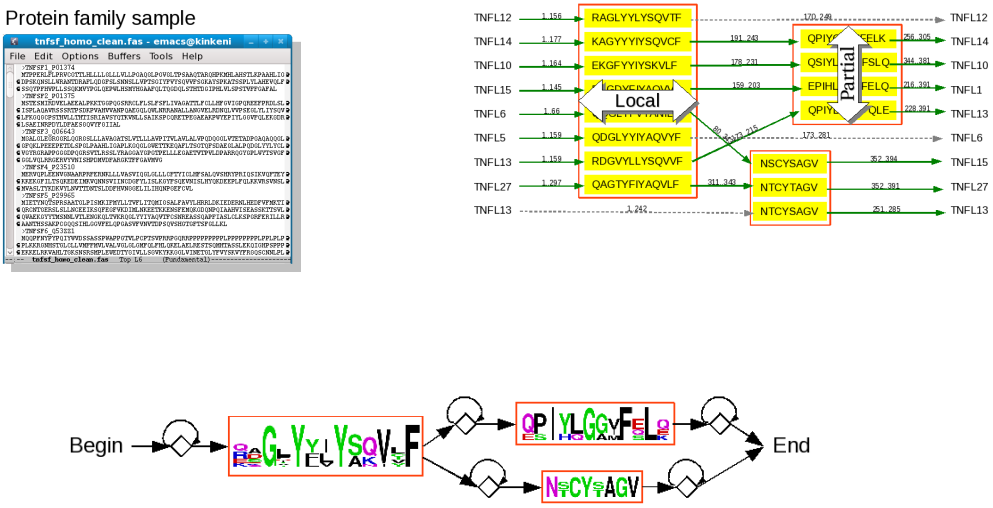


Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences (up left), a partial local alignment is computed (up right) and an automaton is inferred, which models the family and allows to search for its unknown members (down) .*

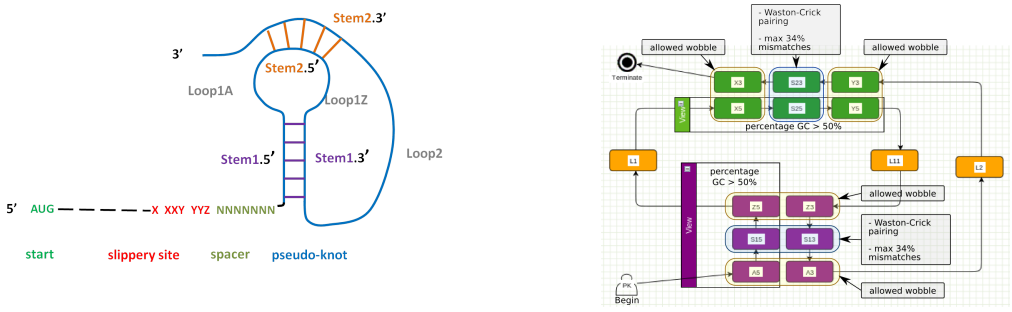


Figure 4. ***Left****: A typical RNA structure: the pseudo-knot.* ***Right****: graphical modeling of a pseudo-knot with String Variable Grammars used in the* ***Logol framework****.*

4.3. 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 (**Axis 2**). 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 Fig. 2). This specific approach is motivated by the quite restricted spectrum of available physiological observations over the asymptotic dynamics of the biological system.

4.4. Biological sequence annotation

In order to check the accuracy of in-silico predictions, a last step (**Axis 3**) 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 figure 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 figure 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.

5. Software

5.1. Data integration: actors involved in the response of a living system

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.

- **Mobyle@GenOuest network portal** We are developing a web service [2] to use several tools to confront knowledge and data towards the correction of large-scale networks, based either on decision diagrams or on answer set programming. BioQuali [3] [4] allows one to confront model and data, localize errors and, when model and data are consistent, to predict the variation of non observed nodes [6]. BioASP [5] was developed in Potsdam and allows one to perform prediction even if model and observations are contradictory, by considering all possible repairs of data and models and computing the common predictions of all repaired models [5]. The portal also include tools for the completion of metabolic networks [32].
- **Combined set of key actors in reaction-based networks: Cadbiom** [6]. This tool is based on state-chart like graphical language. It allows investigating synchronization events in biological networks. It is applied to cancer signaling networks [10].

[2] http://mobyle.genouest.org/cgi-bin/Mobyle/portal.py
[3] http://www.irisa.fr/symbiose/bioquali/
[4] http://www.irisa.fr/symbiose/projects/bioqualiCytoscapePlugin/
[5] http://www.cs.uni-potsdam.de/bioasp/
[6] http://cadbiom.genouest.org/

5.2. Dynamics: actor/parameter combination controlling the response of a system

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

- **caspo: Cell ASP Optimizer**. We have implemented a Python package which combines BioASP [7] and CellNOpt[8] to provide an easy to use software for learning Boolean logic models using ASP [19]. The software is available for download[9] and also as a web service through the Mobyle framework.
- **Event network and quantitative time-series data: POGG** [10]. POGG is a tool developed in collaboration with the LINA lab (Nantes) that uses mean dynamics to score the respective relevance of regulatory pathways in a higher-scale phenotype. It was applied to the quantitative prediction of protein quantities under exponential growth [2]. It predicts the main features of a Markov chain model derived from a reaction-based model when confronted to a single time-series quantitative observation.

5.3. Sequence annotation

We develop tools for discovery and search of complex pattern signatures within biological sequences, with a focus on protein sequences. An integrated environment, Dr Motif [11] is available on the GenOuest Platform that gathers state-of-the-art tools for pattern discovery and pattern matching including our own developments.

- **Complex pattern discovery: Protomata learner** [12] is a grammatical inference framework suitable for the inference of accurate protein signatures [3], [4]. It was completely redesigned in 2010-2011 thanks to a specific Inria action (ADT support). It is currently applied to the recognition of olfactory receptor genes.
- **Complex pattern matching: Logol [13].** We have completely redesigned Stan (suffix-tree analyser), a former tool to search for nucleotidic and peptidic patterns within whole chromosomes [7]. The result is Logol, a software suite accepting a syntax based on String Variable Grammars, which allows the description of realistic complex patterns including ambiguities, insertions/ deletions, gaps, repeats and palindromes. It has been presented for the first time in [21]. Logol has been applied to the detection of -1 frameshifts, a structure including pseudo knots, on a reference benchmark (Recode2).

6. New Results

6.1. Data integration

Participants: Jacques Nicolas [**contact**], Geoffroy Andrieux, Andres Aravena, Pierre Blavy, Jérémie Bourdon, Guillaume Collet, Damien Eveillard, Michel Le Borgne, Sylvain Prigent, Anne Siegel, Sven Thiele, Valentin Wucher.

- **Identification of key regulators by the integration of flux and regulatory information** [*P. Blavy, A. Siegel*] We introduced a new method to combine reaction-based "flux" information (consumption and prediction of molecules) and regulatory "causal" information (effect of the variation of a molecule on the variation of another molecule) in order to find potential key regulators of a set of molecules. It has been validated by recovering among the causal graph derived from the Transpath database the main regulators of 190 groups of genes which are known to share a transcription factor according to the TRED database. [22][Online publication]

[7] http://www.cellnopt.org/
[8] http://www.cs.uni-potsdam.de/ ~sthiele/bioasp/
[9] http://pypi.python.org/pypi/caspo
[10] http://pogg.genouest.org/wiki.php/Home
[11] http://www.drmotifs.org/
[12] http://protomata-learner.genouest.org/
[13] **http://webapps.genouest.org/LogolDesigner/**

- **Reconstruction of transcriptional networks** [*A. Aravena, A. Siegel*] Transcriptional regulatory network models can be reconstructed ab initio from DNA sequence data by locating the binding sites, defined by position specific score matrices, and identifying transcription factors by homology with known ones in other organisms. In general the resulting network contains spurious elements. We use differential expression experimental data, in the form of Mutual Information, as ASP logical constraints to be satisfied by any valid regulatory network subgraph. These rules are used to determine the minimal sets of motif and transcription factors which constitute a genetic regulatory network compatible with experimental data [20][Online publication].
- **Studying diversity in marine environment** [*D. Eveillard*] We proposed a statistical-based data analysis of environmental microarrays. It shows that similar physical parameters drive bacterial and archae communities that share common ammonia oxidizing capacities [12][Online publication]
- **Brown algae metabolic network reconstruction** [*S. Prigent, S. Thiele, A. Siegel*] In order to better understand the functioning of cellular metabolism in the model brown alga *E. siliculosus*, metabolic networks are under construction based on genomic information. Two approaches are conducted in parallel to complete the network, a stochastic one that proceeds by sampling the solution space and a combinatorial one that tries to minimize the number of added reactions [23].

6.2. Asymptotic dynamics

Participants: Anne Siegel [**contact**], Oumarou Abdou-Arbi, Geoffroy Andrieux, Pierre Blavy, Jérémie Bourdon, Damien Eveillard, Michel Le Borgne, Vincent Picard, Sven Thiele, Santiago Videla.

- **Probabilistic sources for sequences and systems biology** [*J. Bourdon*] The habilitation thesis surveys how methods based on average-case analysis of algorithms can be used to model the quantitative response of a biological system from a biomolecular to a physiological scale [28].
- **Learning the early-response of protein signaling networks.** [*S. Videla, S. Thiele, A. Siegel*] We demonstrated the usefulness of the Answer Set Programming approach (ASP) to learn Boolean models from high-throughput phospho-proteomics data. Exact constraint solving showed a quantum leap over heuristic (state-of-the-art) methods in terms of efficiency and scalability, and guarantees global optimality of solutions as well as provides a complete set of solutions [19][Online publication]
- **Numerical model of signaling pathways** [*G. Andrieux, M. Le Borgne*] We have proposed an integrative numerical (ODE) model for the dynamic regulation of $TGF\beta$ Signaling by $TIF1\gamma$. The model successfully unifies the seemingly opposite roles of $TIF1\gamma$, and reveals how changing $TIF1\gamma/Smad4$ ratios affect the cellular response to stimulation by $TGF\beta$, accounting for a highly graded determination of cell fate. [10].
- **Identification of regulatory networks in ecology** [*D. Eveillard*] A clustering data-based approach emphasizes regulatory networks at the bacterial population scale. It allowed the identification of antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats. [15][Online publication]

6.3. Sequence annotation

Participants: François Coste [**contact**], Catherine Belleannée, Gaëlle Garet, Clovis Galiez, Laurent Miclet, Jacques Nicolas.

- **Expressive pattern matching** [*C. Belleannée, J. Nicolas*] We have presented for the first time Logol, a new application designed to achieve pattern matching in possibly large sequences with realistic biological motifs. Logol consists in both a language for describing patterns, and the associated parser for effectively scanning sequences (RNA, DNA or protein) with such motifs. The language, based on an high level grammatical formalism, allows to express flexible patterns (with misparings - improper alignment of DNA strands - and indels) composed of both sequential and structural elements (such as repeats or pseudoknots)[21][Online publication]. Logol has been applied to the detection of -1 frameshifts, a structure including pseudoknots, on a reference benchmark (Recode2) [26][Online publication].

- **Analysis of sequence repeats** [*J. Nicolas*] We have participated to a book that introduces up-to-date methods for the identification and study of transposable elements in genomes. J. Nicolas contributed with a chapter that provides an overview of the formal underpinnings of the search for these highly repeated elements in genomic sequences and describes a selection of practical tools for their analysis. It concludes with the interest of syntactic analysis in this domain [24][Online publication].
- **Grammatical models for local patterns** [*G. Garet, J. Nicolas, F. Coste*] We studied the annotation of new proteins with respect to banks of already annotated protein sequences. For this task, we are developping grammatical inference methods. We introduced new classes of substitutable languages and new generalization criterion based on local substitutability concept and illustrated the great potential of the approach on a benchmark considering a real non trivial protein family. [16][Online publication]
- **Local maximality** [*L. Miclet*] Starting from locally maximal subwords and locally minimal superwords common to a finite set of words, we have defined the corresponding sets of alignments. We gave a partial order relation between such sets of alignments, as well as two operations between them and showed it has a lattice structure that can be used for inducing a generalization of the set of words [18][17].
- **Searching for Smallest Grammars on Large Sequences and Application to DNA** [*F. Coste*] We are motivated by the inference of the structure of genomic sequences, that we address as an instance of the smallest grammar problem. Previously, we reduce it to two independent optimization problems: choosing which words will be constituents of the final grammar and finding a minimal parsing with these constituents. This year we made these ideas applicable on large sequences. First, we improved the complexity of existing algorithms by using the concept of maximal repeats for constituents. Then, we improved the size of the grammars by cautiously adding a minimal parsing optimization step. Together, these approaches enabled us to propose new practical algorithms that return smaller grammars (up to 10%) in approximately the same amount of time than their competitors on a classical set of genomic sequences and on whole genomes. [14] [Online publication].
- **CyanoLyase: a database of phycobilin lyase sequences, motifs and functions** [*F. Coste*] In collaboration with our partners of the ANR project Pelican, we have set up CyanoLyase (http://cyanolyase.genouest.org/), a manually curated sequence and signature database of phycobilin lyases and related proteins. Protomata-Learner has been used to establish the signature of the 32 known subfamilies that are used to rapidly retrieve and annotate lyases from any new genome [13] [Online publication]

7. Partnerships and Cooperations

7.1. Regional Initiatives

7.1.1. Partnership with computer science laboratories in Nantes

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

Methodologies are developed in close collaboration with university of Nantes (LINA) and Ecole centrale Nantes (Irccyn). This is acted through the Biotempo and Idealg ANR projects and co-development of common software toolboxes within the Renabi-GO platform process. Two Ph-D thesis are also co-supervised within these collaborations.

7.1.2. Partnership in Marine Biology

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

A strong application domain of the Dyliss project is marine Biology. This application domain is co-developped with the station biologique de Roscoff and their three UMR and involves several contracts. The IDEALG consortium is a long term project (10 years, ANR Investissement avenir) aiming 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 modelling of the initation of sea-urchin translation (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 P.-F. Pluchon) and the dentification of key actors implied in competition for light in the ocean (PELICAN ANR project).

7.1.3. Partnership with Inra and Health

Participants: Jacques Nicolas, Catherine Belleannée, François Coste, Michel Le Borgne, Anne Siegel, Oumarou Abdou-Arbi, Geoffroy Andrieux, Pierre Blavy, Valentin Wucher.

We have a strong and long term collaboration with biologists of INRA in Rennes : IGEEP and SENAH units. This partnership is acted by the co-supervision of one post-doctorant and two PhD students. It is also reinforced by collaboration within ANR contracts (Lepidolf, MirNadapt, FatInteger).

We also have a strong and long term collaboration with the IRSET laboratory at Univ. Rennes 1, acted by a co-supervised Ph-D thesis. This partnership is reinforced with the ANR contract Biotempo and has been also supported in the framework of the previous CPER by a project, BasicLab, on a lab on chip for cell assays.

7.2. National Initiatives

7.2.1. Long-term contracts

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

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

Cooperation with Univ. of Chile (MATHomics, A. Maass) on methods for the identification of biomarkers and softwares for biochip design. It aims at combining automatic reasoning on biological sequences and networks with probabilistic approaches to manage, explore and integrate large sets of heterogeneous omics data into networks of interactions allowing to produce biomarkers, with a main application to biomining bacteria. Co-funded by Inria and CORFO-chile from 2012 to 2022, the program includes a co-advised ph-D student (A. Aravena) and a post-doc (S. Thiele). In this context, IntegrativeBioChile is an Associate Team between Dyliss and the Laboratory of Bioinformatics and Mathematics of the Genome hosted at Univ. of Chile funded from 2011 to 2013.

7.2.1.2. ANR Idealg

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

IDEALG is one of the five laureates from the national call 2010 for Biotechnology and Bioresource and will run until 2020. It gathers 18 different partners from the academic sector (CNRS, IFREMER, UEB, UBO, UBS, ENSCR, University of Nantes, INRA, AgroCampus), the industrial sector (C-WEED, Bezhin Rosko, Aleor, France Haliotis, DuPont) as well as a technical centre 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.

7.2.2. Methodology: ANR Biotempo

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

The BioTempo projects aims at developing some original methods for studying biological systems. The goal is to introduce partial quantitative information either on time or on component observations to gain in the analysis and interpretation of biological data. Three biological applications are considered regulation systems used by biomining bacteria, TGF$beta$ signaling and initiation of sea-urchin translation. It is funded by ANR Blanc (SIMI2) and coordinated by A. Siegel from 2011 to 2014. [details]

7.2.3. Proof-of-concept on dedicated applications

7.2.3.1. ANR Fatinteger

Participants: Anne Siegel, Jacques Nicolas, Catherine Belleannée, Pierre Blavy.

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

7.2.3.2. ANR Lepidolf

Participants: François Coste, Jacques Nicolas.

The LEPIDOLF project aims at better understanding olfactory mechanisms in insects. The goal is to establish the antennal transcriptome of the cotton leafworm Spodoptera littoralis, a noctuid representative of crop pest insects. It is funded by ANR call Blanc and coordinated by E. Jacquin-Joly from UMR PISC (INRA Versailles) from 2009 to 2012. Our contribution is to use grammatical inference to build characteristic signatures of the Olfactory Receptor family, which will be used to scan directly 454-sequencing reads and available partial cDNAs of genes expressed in the antenna of Lepidoptera or deduced proteins.

7.2.3.3. ANR Mirnadapt

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

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

7.2.3.4. ANR Pelican

Participant: François Coste.

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

7.2.4. Programs funded by research institutions

7.2.4.1. Inria Bioscience Ressource

Participants: Claudia Hériveau, Jacques Nicolas.

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

7.2.4.2. Aquasyst

Participants: Damien Eveillard, Anne Siegel.

PEPS contract 2011-2012 whose goal is to combine Environmental genomics and Systems biology for the understanding of aquifere denitrification.

7.3. European Initiatives

7.3.1. Collaborations with Major European Organizations

Partner: EBI (Great-Britain)

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

Partner: Potsdam university (Germany)

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: SIEGEL Anne

International Partner (Institution - Laboratory - Researcher):

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

Duration: 2011 - 2013

See also: http://www.irisa.fr/symbiose/people/asiegel/EA/

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

7.4.2. Participation In International Programs

7.4.2.1. Argentina - MinCYT-Inria 2011-12

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

Title: Modélisation linguistique de séquences génomiques par apprentissage de grammaires

Financial support: MinCYT-Inria program 2011-12

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

7.4.2.2. International joint supervision of PhD agreement

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

Inria principal investigator: Anne Siegel

International Partners (Institution - Laboratory - Researcher):

University of Chile (Chile)

Duration: Jul 2011 - Jul 2014

Title: Analyse automatisée et générique de réseaux métaboliques en nutrition

Inria principal investigator: Anne Siegel

International Partner (Institution - Laboratory - Researcher):

University of Ouagadougou (Burkina Faso)

Duration: October 2010 - September 2013

7.4.2.3. Germany. Egide Procope Program 2011-12

Program: PHC

Title: Reasoning in systems biology with answer set programming.

Inria principal investigator: Jacques Nicolas

International Partner :

University of Potsdam (Germany)

Institut fur Informatik Wissensverarbeitung und Informationssysteme

T. Schaub

Duration: Jan 2011 - Dec 2012

The cooperation adresses various aspects of the development of the Answer Set Programming approach in bioinformatics.Based on formal methods for the Analysis of big metabolic networks we developed a new approach with Answer Set Programming This approach can be used to check whether a network contains the reaction pathways that explain the bio-synthetic behavior of the organism. Further we developed an approach for the learning of logical models of protein signaling networks.

7.4.2.4. Amadeus (Austria)

Program: PHC

Title: From fractals to numeration

Inria principal investigator: Anne SIEGEL

International Partner (Institution - Laboratory - Researcher):

University of Leoben (Austria)

Duration: Jan 2011 - Dec 2012

7.5. International Research Visitors

7.5.1. Visits of International Scientists

- **Germany**. Department of Computer Science, Potsdam. 5 days [T. Schaub, M. Gebser, M. Ostrowski]
- **Chile**. Centro de Modelimiento Matematico, Santiago. 10 days [A. Maass]

7.5.1.1. Internships

- Internship April to July, 2012. Co-supervised by Anne Siegel and Sylvain Prigent. Student : Floriane Ethys de Corny. Subject: Improvement of metabolic networks. Application to Ectocarpus siliculous.

7.5.2. Visits to International Teams

- **Austria**. Department of Mathematics, Leoben & Vienna. *Dynamical systems*. 5 days [A. Siegel]
- **Burkina-Faso**. Department of Computer Science, Oagadougou. *Multi-objective methods for the static analysis of metabolic network*. 2 months [O. Abdou-Arbi]
- **Chile**. Centro de Modelimiento Matematico, Santiago. *Metabolic modeling of bacteria*. 14 days [D. Eveillard]
- **Chile**. Centro de Modelimiento Matematico, Santiago. *Data integration*. 7 days [A. Siegel]
- **Chile**. Centro de Modelimiento Matematico, Santiago. *Applications of ASP*. 21 days [S. Thiele]
- **Chile**. Centro de Modelimiento Matematico, Santiago. *Applications of ASP*. 10 days [S. Videla]
- **Germany**. Department of Computer Science, Potsdam. *Constraint-based approaches*. 5 days [J. Nicolas]
- **Germany**. Department of Computer Science, Potsdam. *Application of ASP to biology*. 5 days [A. Siegel]
- **Germany**. Department of Computer Science, Potsdam. *Reconstruction of metabolic networks*. 10 days [S. Thiele]
- **Germany**. Department of Computer Science, Potsdam. *Learning logical rules for protein signaling networks*. 2 months [S. Videla]
- **Niger**. University of Maradi. *Multi-objective methods for the static analysis of metabolic network*. 1 month [O. Abdou-Arbi]
- **UK** EMBL-European Bioinformatics Institute. *Learning logical rules for protein signaling networks*. 3 days [S. Videla]

8. Dissemination

8.1. Scientific Animation

8.1.1. Administrative functions: scientific committees, journal boards

- Scientific Advisory Board of ITMO Genetics Genomics and Bioinformatics [J. Nicolas].
- Scientific Advisory Board of GDR BIM " Molecular Bioinformatics"[J. Nicolas].
- Member of the section 01 of the Comité National de la Recherche Scientifique [A. Siegel]
- Member of the IRISA laboratory council [F. Coste]
- Scientific Advisory Board of Biogenouest [J. Bourdon, J. Nicolas, A. Siegel].
- Steering committee of the International Inference community (ICGI) [F. Coste]
- Academic editor: Plos One [J. Bourdon]
- Recruitment committees: junior and senior research (CNRS) [A. Siegel], assistant professor (Irccyn, Univ. Nantes) [A. Siegel, D. Eveillard], professor (Lille) [A. Siegel]
- Referee: IET Systems Biology, Bioinformatics, BMC Bioinformatics, PLoS One, Annals of combinatorics, Medical & Biological Engineering & Computing, Theoretical Computer Science, Algorithmic Learning Theory '12
- Member of SCAS (Service Commun d'Action Sociale) of Univ. Rennes 1 [C. Belleannée]

8.1.2. JOBIM

JOBIM JOBIM is the scientific yearly appointment of the French-speaking bioinformatics community. The official languages of the conference are English and French. It has been organized in Rennes this year from 3 to 6 July 2012 (P. Peterlongo, C. Lemaitre and teams Genscale and Dyliss) and chaired by F. Coste and D. Tagu (Inra). Invited speakers were David B. Searls, University of Pennsylvania, Hugues Roest Crollius, Ecole Normale Supérieure, Pierre Baldi, University of California in Irvine, Ivo Hofacker, Institute for Theoretical Chemistry, Toni Gabaldon, Centre for Genomic Regulation, Martin Vingron, Max Planck Institute for Molecular Genetics and Bertil Schmidt, Johannes Gutenberg University Mainz. The program contained 33 papers (acceptance rate 50%) grouped into ten scientific sessions including sequence analysis, the study of protein interactions, and works on regulation and evolution.

[25] [Online publication]

8.1.3. Ecole Jeunes Chercheurs en Informatique-Mathématiques (GDR IM)

The GDR Informatique Mathématique has organized a young researcher seminar from 19 to 23 March in Rennes (Chair: A. Siegel. Organization M.-N. Georgeault and E. Lebret) http://ejcim2012.irisa.fr. It gathered 93 participants and proposed several courses including biological sequence linguistics (J. Nicolas and F. Coste), time modelling for dynamic systemchecking (D. Eveillard) and practical applications of discrete dynamic systems (J. Bourdon and A. Siegel).

8.1.4. Local meetings

- **IDEALG annual meeting** The IDEALG annual meeting has been organized in Irisa/Inria Rennes from 27 to 28 September 2012. Dyliss is participating to bioinformatics studies for the development of an integrated platform on seaweed omics data.
- **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. [web site].

8.1.5. Conference program committees

- Cap'2012 [F. Coste]
- JOBIM [F. Coste/co-head, J. Bourdon, J. Nicolas, A. Siegel]
- ICGI'12 [F. Coste]
- Numeration [A. Siegel]

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

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

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

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

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

Licence : G. Garet, Office automation, 20h, L1, Univ. Rennes 1, France.

Licence : G. Garet, Functional algorithm, 24h, L1, Univ. Rennes 1, France.

Licence : G. Garet, Programming, 22h, L3, Univ. Rennes 1, France.

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

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

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

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

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

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

Master: V. Picard, Préparation à l'agrégation de mathématiques option D: épreuve de modélisation, 12h, L2, ENS Rennes/Univ. Rennes 1, France

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

Master: C. Belleannée, Architecture matérielle et interface au système, 28h, M2 informatique, Rennes1 France

Master: F. Coste, Apprentissage Supervisé, 15h, M2 Informatique, Univ. Rennes 1, France

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

Doctorat : J. Bourdon, Applications de systèmes dynamiques discrets, 2h, Ecole Jeunes Chercheurs en Informatique Mathématique, Rennes, France

Doctorat : J. Bourdon, Réseaux biologiques semi-quantitatifs : dynamique et propriétés émergentes , 2h, Ecole thématique OSUR sur les systèmes complexes, Rennes, France

Doctorat : F. Coste & J. Nicolas, Linguistique des séquences biologiques, 4h, Ecole Jeunes Chercheurs en Informatique Mathématique, Rennes, France

Doctorat: D. Eveillard, Modélisation du temps pour la vérification des systèmes dynamiques, 2h, Ecole Jeunes Chercheurs en informatique mathématique, Rennes, France.

Doctorat: D. Eveillard, From Omics data to models, 4h, Ecole Thématique Ecologie et Génomique Environnementale, Aussois, France.

Doctorat : A. Siegel, Introduction aux systèmes dynamiques, 2h, Ecole Jeunes Chercheurs en Informatique Mathématique, Rennes, France

8.2.2. Seminars

- J. Bourdon , *Quelques outils pour étudier la dynamique des réseaux génétiques*, 10th days of the GenOuest platform: biological networks, 2012.
- F. Coste, *Characterization of protein families: overpassing HMM expressivity*, second meeting Idealg, Rennes sept. 2012
- D. Eveillard, *Temporal and quantitative behaviors of biological systems - Can we learn something by modeling via a systems biology viewpoint ?*, séminaire de la station biologique de Roscoff, Roscoff.
- D. Eveillard, *Quantitative modeling of biological systems*, Biocore seminar, Inria Sophia-Antipolis.
- G. Andrieux, *Analyzing Large Models of TGFbeta with Cadbiom and the Process Hitting*, Ecole Jeunes Chercheurs en Informatique Mathématique 2012, Rennes.
- S. Prigent, *Methods of metabolic network reconstruction: corresponding contributions*, second meeting Idealg, Rennes sept. 2012
- S. Prigent, *Reconstruction de réseaux métaboliques par la programmation logique*, Ecole Jeunes Chercheurs en Informatique Mathématique 2012, Rennes.
- S. Prigent, *Que nous apprend la reconstruction d'un réseau métabolique ?*, 10th days of the genouest platform: biological networks, 2012.
- A. Siegel, *A review on Pisot conjecture, coincidence conditions and related graphs*, TU Wien, department of mathematics, 2012.
- A. Siegel, *Using contraints programming to investigate the robustness of biological networks reconstruction*, University of Chile, 2012.
- A. Siegel, *Quelques approches formelles pour tester la robustesse de processus de reconstruction de réseaux*, Séminaire du réseau NetBio, 2012.
- V. Wucher. *Modélisation d'un réseau de régulation d'ARN pour prédire des fonctions de gènes impliqués dans le mode de reproduction du puceron du pois*. Journées "Bioinformatique des ARNnc", Toulouse.

8.2.3. Supervision

HdR : Jérémie Bourdon, *Sources probabilistes : des séquences aux systèmes*, Université de Nantes, 5 décembre 2012 [28].

PhD in progress : Oumarou Abdou-Arbi *Analyse Automatisée et générique des réseaux métaboliques en nutrition*, started in October 2010, supervised by A. Siegel and T. Tabsoba (Burkina-Faso).

PhD in progress : Geoffroy Andrieux, *Discrete approach modeling of biological signaling pathway*, started in October 2010, supervised by N. Théret (Inserm) and M. Le Borgne

PhD in progress : Andres Aravena, *Introduire des approches combinatoires dans des modèles probabilistes pour la découverte d'évènements de régulation d'un système biologique à partir de données hétérogènes*, started in July 2011, supervised by A. Maass (CMM, University of Chile) and A. Siegel.

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

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

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 September 2012, supervised by A. Siegel and J. Bourdon.

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

PhD in progress : Santiago Videla, *Applying logic programming to the construction of robust predictive and multi-scale models of bioleaching bacteria*, started in November 2011, supervised by A. Siegel

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

8.2.4. Juries

- *Member of habilitation thesis jury*. J. Bourdon, Université de Nantes [A. Siegel].
- *Member of Ph-D thesis jury*. P. Vanier, Université de Marseille [A. Siegel]. M. Noual, ENS Lyon [A. Siegel]. P. Bordron, Univ. Nantes [D. Eveillard]
- *Referee of Ph-D thesis*. S. Thiele, University of Potsdam [A. Siegel]. N. Loira, University of Bordeaux [A. Siegel].

9. Bibliography

Major publications by the team in recent years

[1] T. BAUMURATOVA, D. SURDEZ, B. DELYON, G. STOLL, O. DELATTRE, O. RADULESCU, A. SIEGEL. *Localizing potentially active post-transcriptional regulations in the Ewing's sarcoma gene regulatory network.*, in "BMC Systems Biology", 2010, vol. 4, n^{o} 1, 146 [*DOI :* 10.1186/1752-0509-4-146], http://hal.inria.fr/inria-00538133.

[2] J. BOURDON, D. EVEILLARD, A. SIEGEL. *Integrating quantitative knowledge into a qualitative gene regulatory network.*, in "PLoS Computational Biology", September 2011, vol. 7, n^{o} 9, e1002157 [*DOI :* 10.1371/JOURNAL.PCBI.1002157], http://hal.archives-ouvertes.fr/hal-00626708.

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

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

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

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

[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, p. 4408-4410, http://hal.archives-ouvertes.fr/hal-00015234.

[8] A. SIEGEL, O. RADULESCU, M. LE BORGNE, P. VEBER, J. OUY, S. LAGARRIGUE. *Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks*, in "Biosystems", 2006, vol. 84, p. 153-174, http://hal.inria.fr/inria-00178809.

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

Publications of the year

Articles in International Peer-Reviewed Journals

[10] G. ANDRIEUX, N. THERET, M. LE BORGNE, L. FATTET, R. RIMOKH. *Dynamic Regulation of Tgf-B Signaling by Tif1γ: A Computational Approach*, in "PLoS ONE", March 2012, http://hal.inria.fr/hal-00760066.

[11] V. BERTHÉ, T. JOLIVET, A. SIEGEL. *Substitutive Arnoux-Rauzy sequences have pure discrete spectrum*, in "Uniform Distribution Theory", 2012, vol. 7, n^{o} 1, p. 173-197, http://hal.inria.fr/hal-00750209.

[12] N. J. BOUSKILL, D. EVEILLARD, D. CHIEN, A. JAYAKUMAR, B. B. WARD. *Environmental factors determining ammonia-oxidizing organism distribution and diversity in marine environments.*, in "Environmental Microbiology", March 2012, vol. 14, n^{o} 3, p. 714-29 [*DOI :* 10.1111/J.1462-2920.2011.02623.X], http://hal.inria.fr/hal-00661746.

[13] 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, p. D396-D401 [*DOI :* 10.1093/NAR/GKS1091], http://hal.inria.fr/hal-00760946.

[14] R. CARRASCOSA, F. COSTE, M. GALLÉ, G. INFANTE-LOPEZ. *Searching for Smallest Grammars on Large Sequences and Application to DNA*, in "Journal of Discrete Algorithms", February 2012, vol. 11, p. 62-72 [*DOI :* 10.1016/J.JDA.2011.04.006], http://hal.inria.fr/inria-00536633.

[15] R. A. LONG, D. EVEILLARD, S. L. M. FRANCO, E. REEVES, J. L. PINCKNEY. *Antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats.*, in "FEMS Microbiology Ecology", August 2012, vol. 83, p. 74-81 [*DOI :* 10.1111/J.1574-6941.2012.01457.X], http://hal.inria.fr/hal-00752749.

International Conferences with Proceedings

[16] F. COSTE, G. GARET, J. NICOLAS. *Local Substitutability for Sequence Generalization*, in "ICGI 2012", Washington, United States, J. HEINZ, C. DE LA HIGUERA, T. OATES (editors), JMLR Workshop and Conference Proceedings, MIT Press, September 2012, vol. 21, p. 97-111, http://hal.inria.fr/hal-00730553.

[17] L. MICLET, N. BARBOT, B. JEUDY. *A Lattice of Sets of Alignments Built on the Common Subword in a Finite Language*, in "International Conference on Grammatical Inference", United States, 2012, vol. 21, p. 164-176, http://hal.inria.fr/hal-00726166.

[18] 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 "1st International Workshop on Similarity and Analogy-based methods in AI (SAMAI)", Montpellier, France, G. RICHARD (editor), August 2012, vol. Rapport interne IRIT/RR-2012-20-FR, p. 25-31, http://hal.inria.fr/hal-00760722.

[19] *Best Paper*
S. VIDELA, C. GUZIOLOWSKI, F. EDUATI, S. THIELE, N. GRABE, J. SAEZ-RODRIGUEZ, A. SIEGEL. *Revisiting the Training of Logic Models of Protein Signaling Networks with a Formal Approach based on Answer Set Programming*, in "CMSB - 10th Computational Methods in Systems Biology 2012", London, United Kingdom, D. GILBERT, M. HEINER (editors), Springer, 2012, vol. 7605, p. 342-361 [*DOI :* 10.1007/978-3-642-33636-2_20], http://hal.inria.fr/hal-00737112.

National Conferences with Proceeding

[20] A. ARAVENA, C. GUZIOLOWSKI, A. SIEGEL, A. MAASS. *Using Mutual Information and Answer Set Programming to refine PWM based transcription regulation network*, in "JOBIM 2012, 13e Journées Ouvertes en Biologie, Informatique et Mathématiques", Rennes, France, July 2012, 171, http://hal.inria.fr/hal-00740722.

[21] C. BELLEANNÉE, O. SALLOU, J. NICOLAS. *Expressive Pattern Matching with Logol. Application to the Modelling of -1 Ribosomal Frameshift events*, in "JOBIM 2012, 13e Journées Ouvertes en Biologie, Informatique et Mathématiques", Rennes, France, July 2012, p. 5-14, http://hal.inria.fr/hal-00726791.

[22] P. BLAVY, F. GONDRET, S. LAGARRIGUE, J. VAN MILGEN, A. SIEGEL. *Using large scale knowledge database about reactions and regulations to find key regulators of sets of genes*, in "JOBIM 2012, 13e Journées Ouvertes en Biologie, Informatique et Mathématiques", Rennes, France, F. COSTE, D. TAGU (editors), Inria Rennes Bretagne Atlantique, 2012, p. 123-132, http://hal.inria.fr/hal-00750512.

Conferences without Proceedings

[23] T. TONON, P. BONIN, S. PRIGENT, Z. SHAO, A. GROISILLIER, S. ROUSVOAL, S. GOULITQUER, J. BOURDON, D. EVEILLARD, C. BOYEN, A. SIEGEL. *Systems biology approaches at cellular level in the model organism Ectocarpus siliculosus to better understand brown algal physiology*, in "Esil 2012: algal post-genomics", Roscoff, France, April 2012, http://hal.inria.fr/hal-00760640.

Scientific Books (or Scientific Book chapters)

[24] J. NICOLAS. *To detect and analyze sequence repeats whatever be their origin*, in "Mobile genetic elements: protocols and genomic applications", Y. BIGOT (editor), Springer Protocols, Springer, February 2012, vol. 859, p. 69-90 [*DOI :* 10.1007/978-1-61779-603-6], http://hal.inria.fr/hal-00730207.

Books or Proceedings Editing

[25] F. COSTE, D. TAGU (editors). *JOBIM 2012, 13e Journées Ouvertes en Biologie, Informatique et Mathématiques, 3-6 Juillet 2012, Rennes*, Inria Rennes - Bretagne Atlantique, June 2012, 558, http://hal.inria.fr/hal-00740287.

Research Reports

[26] A. ROCHETEAU, C. BELLEANNÉE. *Recherche d'éléments structurés dans les génomes par modèles logiques*, Inria, June 2012, n^{o} PI-1994, 30, http://hal.inria.fr/hal-00684388.

References in notes

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

[28] J. BOURDON. *Sources probabilistes: des séquences aux systèmes*, Université de Nantes, Nantes, December 2012, 115 pages, Document dH´abilitation à Diriger des Recherches.

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

[30] 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.

[31] 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, p. 226-236 [*DOI :* 10.1093/BIOINFORMATICS/17.3.226], http://bioinformatics.oxfordjournals.org/content/17/3/226.abstract.

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

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

[34] 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, p. 73-102.