IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Activity Report 2015

Project-Team DYLISS

Dynamics, Logics and Inference for biological Systems and Sequences

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

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Computational Biology

Table of contents

Project-Team DYLISS

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

Keywords:

Computer Science and Digital Science:

3.2.4. - Semantic Web
3.2.5. - Ontologies
3.3. - Data and knowledge analysis
7.2. - Discrete mathematics, combinatorics
7.3. - Operations research, optimization, game theory
7.4. - Logic in Computer Science
8.1. - Knowledge
8.2. - Machine learning
8.7. - AI algorithmics

Other Research Topics and Application Domains:

1.1.11. - Systems biology
1.1.2. - Molecular biology
1.1.3. - Cellular biology
1.1.9. - Bioinformatics
2.2.3. - Cancer

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

Engineers

Charles Bettembourg [CNRS, until Dec 2015, granted by ANR MiRNAdapt project]
Marie Chevallier [Inria]
Guillaume Collet [CNRS, until Mar 2015, granted by ANR IDEALG project]
Jeanne Got [CNRS]
Yann Guitton [CNRS, until May 2015, granted by ANR IDEALG project]
Camille Trottier [CNRS, from Oct 2015, granted by ANR IDEALG project]
Meziane Aite [Inria, from Nov 2015]

PhD Students

Aymeric Antoine-Lorquin [Univ. Rennes I]
Jean Coquet [Univ. Rennes I]
Victorien Delannée [Univ. Rennes I]
Clémence Frioux [Inria, from Feb 2015]
Clovis Galiez [Inria]

Julie Laniau [Inria]
Vincent Picard [Univ. Rennes I, until Aug 2015]

Post-Doctoral Fellow
Aurélie Evrard [INRA, from Jul 2015]

Visiting Scientists
Oumarou Abdou Arbi [Assistant prof, from Jan 2015 until Feb 2015]
Mauricio Latorre [Assistant Prof, Chile, from Sep 2015 until Oct 2015]

Administrative Assistant
Marie Le Roic [Univ. Rennes I]

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 bioinformatics Dyliss team is sequence analysis and systems biology. Our main goal in biology is to characterize groups of genetic actors that control the phenotypic answer of non-model species when challenged by their environment. Unlike model species, only 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.

At the sequence level, the main challenge is to transfer information available in genomes of 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) [6].

Solving these computational issues requires addressing NP-hard qualitative (non-temporal) issues. We have developed a long-term collaboration with Potsdam University in order to use a logical paradigm named **Answer Set Programming**(ASP) [43], [66] 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 [45], and proposed robust corrections [5]. See Fig. 1 for details. The results obtained so far suggest that this approach is compatible with efficiency, scale and expressivity needed by biological systems. Our goal is now to provide **formal models of queries on biological networks** with the focus of integrating dynamical information as explicit logical constraints in the modeling process. This would definitely introduce such logical paradigms as a powerful approach to build and query reconstructed biological systems, in complement to discriminative approaches. Notice that our main issue is in the field of knowledge representation. More precisely, we do not wish to develop new solvers or grounders, a self-contained computational issue which is addressed by specialized teams such as our collaborator team in Potsdam. Our goal is rather to investigate whether progresses in the field of constraint logical programming, shown by the performance of ASP-solvers in several recent competitions, are now sufficient to address the complexity of constraint-satisfiability and combinatorial optimization issues explored in systems biology.

By exploring the complete space of models, our approach typically produces numerous candidate models compatible with the observations. We began investigating to what extent domain knowledge can further refine the analysis of the set of models by identifying classes of similar models, or by selecting the models that best fit biological knowledge. We anticipate that this will be particularly relevant when studying non-model species for which little is known but valuable information from other species can be transposed or adapted. These efforts consist in developing reasoning methods based on ontologies as formal representation of symbolic knowledge.

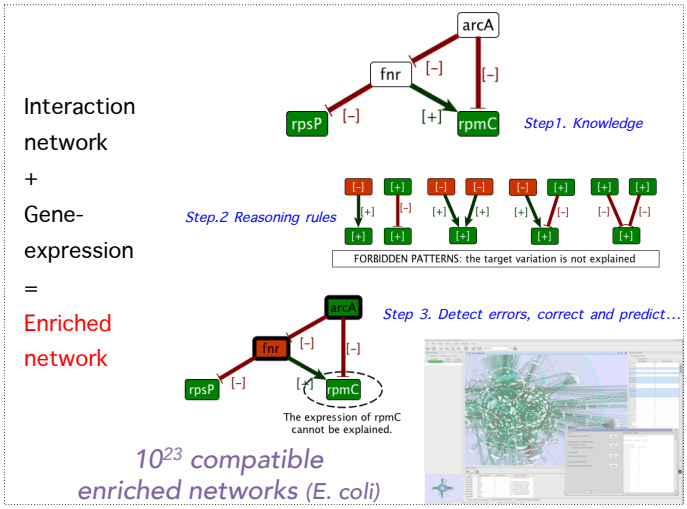

Figure 1. An example of reasoning process in order to identify which expression of non-observed nodes (white nodes) are fixed by partial observations and rules derived from the system dynamics. [6], [5] ***Step 1****. Regulation knowledge is represented as a signed oriented graph. Edge colors stand for regulatory effects (red/green→ inhibition or activation). Vertex colors stand for known 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).*

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 optimization techniques to learn models of the dynamics of a biology systems compatible with a set of quantitative measurements in order to predict its quantitative response at a larger-scale. Our framework mixes mechanistic and probabilistic modeling [2]. The system is modeled by an Event Transition Graph, that is, a **Markovian qualitative description of its dynamics** together with quantitative laws which describe the effect of the dynamic transitions over higher scale quantitative measurements. Then, a few time-series quantitative 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 [65]. This property can be derived into constraints to describe the set of admissible weighted Markov chains whose asymptotic behavior agrees with the quantitative measures at hand. A precise study of this constrained space via local search optimization emphasizes the most important discrete events that must occur to reproduce the information at hand. These methods have been validated on the *E. coli* regulatory network benchmark. See Figure 2 for illustration. We now plan to apply these techniques to reduced networks representing the main pathways and actors automatically generated from the integrative methods developed in the former section. This requires to improve the range of dynamics that can be modeled by these techniques, as well as the efficiency and scalability of the local search algorithms.

3.3. Modeling sequences with formal grammars

Our research on modeling biomolecular sequences with expressive formal grammars focuses on learning such grammars from examples, helping biologists to design their own grammar and providing practical parsing tools.

On the development of machine learning algorithms for the induction of grammatical models [33], we have a strong expertise on learning finite state automata. 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 tool allowed us to properly model the TNF protein family, a difficult task for classical probabilistic-based approaches (see Fig. 3). It was also applied successfully to model important enzymatic families of proteins in cyanobacteria [3]. Our future goal is to further demonstrate the relevance of formal language modeling by addressing the question of a fully automatic prediction from the sequence of all the enzymatic families, aiming at improving even more the sensitivity and specificity of the models. As enzyme-substrate interactions are very specific central relations for integrated genome/metabolome studies and are characterized by faint signatures, we shall rely on models for active sites involved in cellular regulation or catalysis mechanisms. This requires to build models gathering both structural and sequence information 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.

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 [70]. An extended formalism,

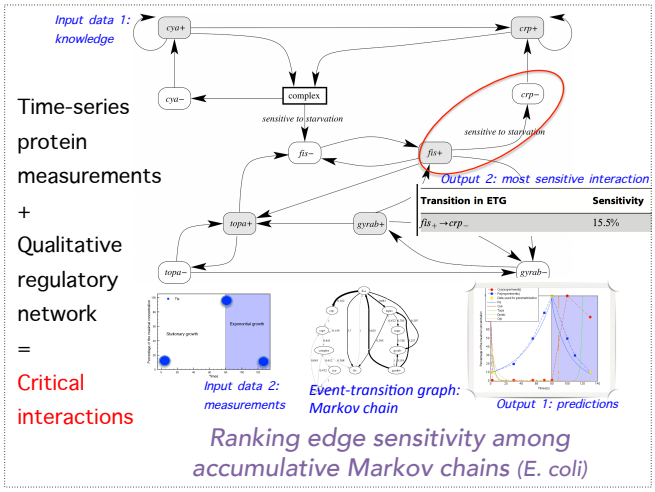

*Figure 2. Prediction of the quantitative behavior of a system using average-case analysis of dynamical systems and identification of key interactions [2]. **Input data** include a qualitative description of the system dynamics at the transcription level (interaction graph) and 3 concentration measurements 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. Local searches 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 best fits with observations. Additionally, literature evidences that high sensitivity ETG transitions correspond to key interaction in E. Coli response to nutritional stress.*

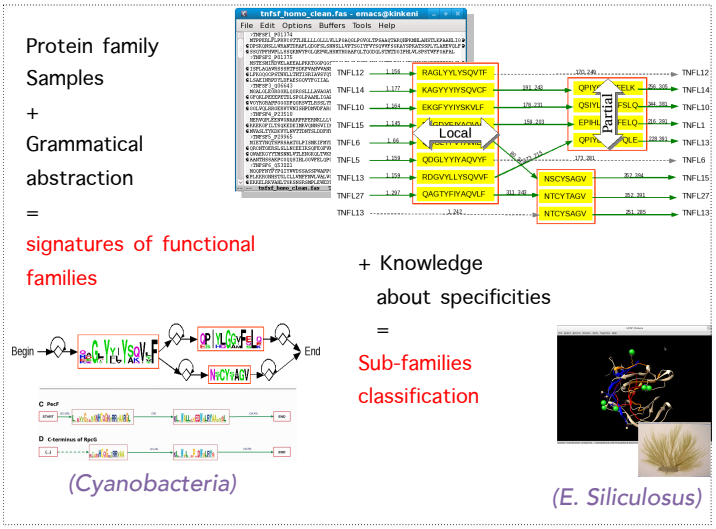

Figure 3. ***Protomata Learner workflow****. Starting from a set of protein sequences, a partial local alignment is computed and an automaton is inferred, which can be considered as a signature of the family of proteins. This allows searching for new members of the family [3]. Adding further information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

String Variable Grammars (SVGs), introduces variables that can be associated to a string during a pattern search (see Fig. 4) [85], [84]. This increases the expressivity of the formalism towards mildly context sensitive grammars. Thus, those grammars model not only DNA/RNA sequence features but also structural features such as repeats, palindromes, stem/loop or pseudo-knots. We have designed a first tool, STAN (suffix-tree analyser), in order to make it possible to search for a subset of SVG patterns in full chromosome sequences [8]. This tool was used for the recognition of transposable elements in *Arabidopsis thaliana* [88] or for the design of a CRISPR database [10]. See Figure 4 for illustration. Our goal is now 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 [1], a language and framework based on a systematic introduction of constraints on string variables.

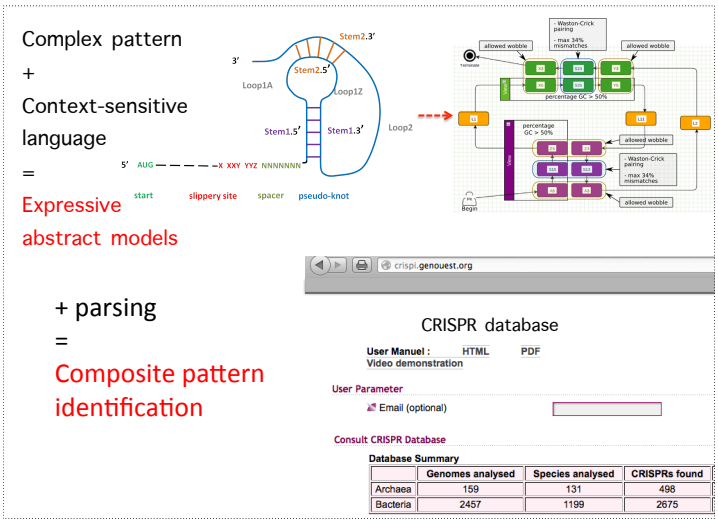

Figure 4. Graphical modeling of a pseudo-knot (RNA structure) based on the expressivity of String Variable Grammars used in the Logol framework. Combined with parsers, this leads to composite pattern identification such as CRISPR [79].

3.4. Symbolic methods for model space exploration: Ontologies and Formal Concepts Analysis

All methods presented in the previous section usually result in pools of candidates which equivalently explain the data and knowlegde. These candidates can be dynamical systems, compounds, biological sequences, proteins... In any case, the output of our formal methods generally requires a posteriori investigation and filtering. We rely on two classes of symbolic technics to this end: Semantic Web technologies and Formal Concept Analysis (FCA). They both aim at the formalization and management of knowledge, that is, the explicitation of relations occuring in structered data. These technics are complementary: The production of relevant concepts in FCA highly depends on the availability of semantic annotations using a controlled set of terms and conversely, building ontologies is a complex process that can be made much easier with FCA.

3.4.1. Semantic web for life sciences

Life sciences are intrinsically complicated and complex. Until a few years ago, both the scarcity of available information and the limited processing power imposed the double constraints that work had to be performed on fragmented areas (either precise but narrow or broad but shallow) as well as using simplifying hypotheses [52]. The recent joint evolution of data acquisition capabilities in the biomedical field, and of the methods and infrastructures supporting data analysis (grids, the Internet...) resulted in an explosion of data production in complementary domains (*omics, phenotypes and traits, pathologies, micro and macro environment...) [52], [56], [47]. This "data deluge" is the life-science version of the more general "big data" phenomenon, with the specificities that the proportion of generated data is much higher, and that these data are highly connected [86]. In addition to the breakthrough in each of these domains, major efforts have been undertaken notably in Systems Biology for developing the links between them. **The bottleneck that once was data scarcity now lies in the lack of adequate methods supporting data integration, processing and analysis**. Each of these steps typically hinges on domain knowledge, which is why it resists automation. This knowledge can be seen as the set of rules representing in what conditions data can be used or can be combined for inferring new data or new links between data.

The knowledge we are focusing on is mostly symbolic, as opposed to other kinds of biomedical knowledge (probabilistic, related to chemical kinetics, 3D models of anatomical entities or 4D models of processes...). It should typically support generalization, association and deduction. There is a long tradition of works in order to come up with an explicit and formal representation of this knowledge that would support automatic processing.

This line of work resulted in the now widespread acceptance of ontologies [87], [59] to represent the biomedical entities, their properties and the relations between these entities. Bard et al. defined **ontologies** as "formal representations of knowledge in which the essential terms are combined with structuring rules that describe the relationships between the terms" [44]. Ontologies range from fairly simple hierarchies to semantically-rich organization supporting complex reasoning [59]. Ontologies are now a well established field [59], [54] that evolved from concept representation [83].

The emergence of ontologies in biomedical informatics and bioinformatics happened in parallel with the development of the **Semantic Web** in the computer science community [81], [83]. The Semantic Web is an extension of the current Web that provides an infrastructure integrating data and ontologies in order to support unified reasoning.

Life sciences are a great application domain for the Semantic Web [57], [76], [46]. Semantic Web technologies have become an integral part of translational medicine and translational bioinformatics [47], [58]. The Linked Data initiative [51] and particularly the Linked Open Data project promotes the integration of data sources in machine-processable formats compatible with the Semantic Web. Figure 5 shows the importance of life sciences. In the past few years, this proved instrumental for addressing the problem of data integration [68], [72].

We are working on the integration of Semantic Web resources with our data analysis methods in order to take existing biological knowledge into account.

3.4.2. Formal Concept Analysis

Initially developed in the community of set and order theorists, algebraists and discrete mathematicians, formal concept analysis aims at the development of conceptual structures which can be logically activated for the formation of judgments and conclusions [90]. In its most simple form, one considers a binary relation between a set of objects O and a set of attributes A. The derivation operator ' associates to each subset U of O (resp. V of A) the subset of elements in A (resp. O) related to all elements in U (resp. V). A formal concept is characterized by an extension (subset of O, individuals belonging to the concept) and an intension (subset of A, properties applying to all objects in the extension), such that the two subsets are stable sets under the double derivation relation ". Concepts are related within a lattice structure (Galois connection) by subconcept-superconcept relations, and this allows to draw causality relations between attribute subsets.

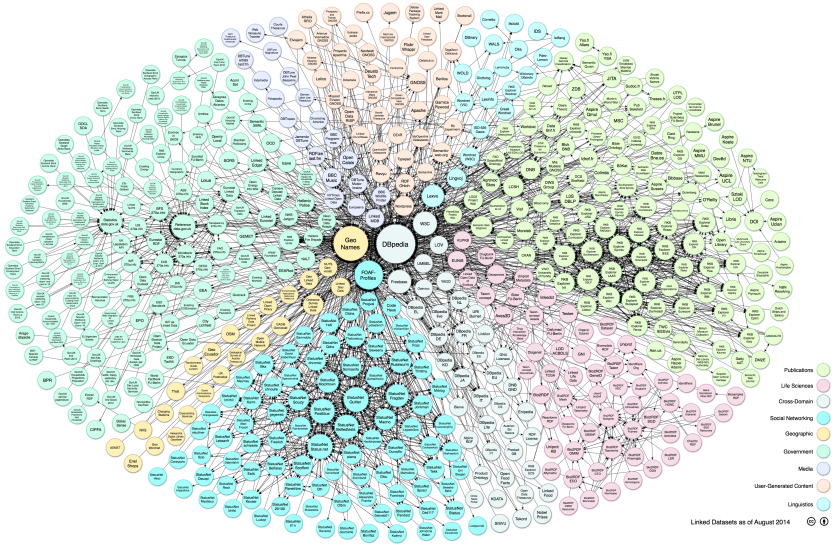

Figure 5. Linked Open Data cloud in August 2014. Nodes are resources. Edges are cross-references between resources. Life science resources constitute the purple portion in the lower right corner.

It is used in various domains managing structered data such as knowledge processing, information retrieval or classification [73]. We study the issues raised by its application in bioinformatics. Among others, it has been used to derive phylogenetic relations among groups of organisms [71], a classification task that requires to take into account many-valued Galois connections. We have proposed in a similar way a classification scheme for the problem of protein assignment in a set of protein families [61]. One of the most important issue with concept analysis is due to the fact that current methods remain very sensitive to the presence of uncertainty or incompleteness in data. On the other hand, this apparent defect can be reversed to serve as a marker of incompleteness or inconsistency. This has been used for example for the drug repositioning issue [62], where the completion of concepts is used as a support for the prediction of new relations in a drug-target-disease network and ultimately the assignment of drugs to new diseases. We have proposed a methodology to tackle the problem of uncertainty on biological networks where edges are mostly predicted links with a high level of false positives [91]. The general idea consists to look for a tradeoff between the simplicity of the conceptual representation and the need to manage exceptions. We are also interested in using ontologies to help this process or to help ontology refinement using concept analysis [74], [50], [78].

Networks are widely used in bioinformatics for the integration of multiple sources of data inside a common model and this leads to very large networks (protein/protein interactions, signaling or regulation network, metabolic network...). Common difficult tasks in this context are visualization, search for local structures (graph mining) and network comparison. Network compression is a good solution for an efficient treatment of all these tasks. This has been used with success in power graphs, which are abstract graphs where nodes are clusters of nodes in the initial graph and edges represent bicliques between two sets of nodes [80]. In fact, concepts are maximal bicliques and we are interested in developing the power graph idea in the framework of concept analysis.

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.

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

4.2. Application fields

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 [77], [64]. 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 [61]. The same tool also allowed us to gain a better understanding of cyanobacteria proteins [3]. At the larger level of 4D structures, classification technics have also allowed us to introduce new methods for the characterization of viruses in marine metagenomic sample [18]. Finally, in dynamical systems, we use asymptotic analysis (tool *pogg*) to decipher the initiation of sea urchin translation [49]. We are currently in two new applications in this domain: the team participates to a Inria Project Lab program with the Biocore and Ange Inria teams, focused on the understanding on green micro-algae; and we are involved in the deciphering of phytoplancton variability at the system biology level in collaboration with the Station Biologique de Roscoff (ANR Samosa).

In **micro-biology**, our main issue is the understanding of bacteria living in extreme environments, mainly in collaboration with the group of bioinformatics at Universidad de Chile (funded by CMM, CRG and Inria-Chile). In order to elucidate the main characteristics of these bacteria, we develop efficient methods to identify the main groups of regulators for their specific response in their living environment. To that purpose, we use constraints-based modeling and combinatorial optimization. The integrative biology tools *meneco bioquali, ingranalysis, shogen, lombarde* were designed in this context [6]. In parallel, in collaboration with Ifremer (Brest), we have conducted similar work to decipher protein-protein interactions within archebacteria [75]. Our sequence analysis tool (*logol*) allowed us to build and maintain a very expressive CRISPR database [10] [48].

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 [40] and characterize upstream transcriptional regulators for such a response [53], with relevant applications in porks [24], [37]. The pattern matching tool *logol* also allows for a fine identification of transcription factor motifs [63] [48]. Constraints-based programming also allows us to decipher regulators of reproduction for pea aphids [69], [92]. Semantic-based analysis was useful for interpreting differences of gene expression in pork meat [67].

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 [45]. 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 [89] and N. Theret group at Inserm/Irset (Rennes) [42]. The dynamical system tools *caspo and cadbiom* were designed within these collaborations. Ongoing studies focus on the understanding of the metabolism of xenobiotics (mecagenotox program) and the filtering of sets of regulatory compounds within large-scale signaling network (TGFSysBio project).

5. Highlights of the Year

5.1. Highlights of the Year

The main novelty in 2015 was the use of Semantic Web technologies to support the integration and query and investigation of large-scale heterogeneous databases. These technologies were applied in the framework of the MiRNAdapt project (funded by ANR) to design a tool for representing and querying bio-molecular information. The tool Askomics was designed in this perspective. In addition, Semantic Web technologies are currently combined with Formal Concept Analysis, to decipher the main regulators of complex systems, with application in cancer system biology (novel project funded by Plan Cancer).

6. New Software and Platforms

6.1. AskOmics

KEYWORDS: Bioinformatics - Linked data - Networks - Semantic Web - Omics
FUNCTIONAL DESCRIPTION

This tool was designed in 2015 in the framework of the MIRNadapt project. Biological studies and bioinformatical analysis produce numerous heterogeneous data, calling for their integration. AskOmics is an integration and interrogation software relying on an RDF model and the SPARQL query language. Its purpose is to obtain quick answers to biological questions demanding currently hours of manual search in several spreadsheet results files. New study perpectives will arise from these answers and from this integration work. Using AskOmics, we integrated an omic dataset borrowed from the MiRNAdapt ANR project that aims to describe the networks of the genes involved in aphids adaptation to seasons. AskOmics allows biologists to integrate and interrogate themselves their data without needing any knowledge about RDF and SPARQL. The query process consists in linking sets of biological entities as nodes in a graphical interface, optionaly specifying biological attributes for these nodes. The graph is then converted inton a SPARQL query to provide the user an answer to his biological questions. The answers are the elements of the sets that match the query contraints.

- Participants: Charles Bettembourg, Anthony Bretaudeau, Olivier Dameron, Aurélie Evrard, Yvanne Chaussin, Anne Siegel, Fabrice Legeai
- Partners: INRA IGEPP
- Contact: Fabrice Legeai
- URL: http://bipaa.genouest.org/askomics_aphid/

6.2. VIRALpro

FUNCTIONAL DESCRIPTION

VIRALpro is a predictor capable of identifying capsid and tail protein sequences using support vector machines (SVM) with an estimated accuracy between 90% and 97%. Predictions are based on the protein amino acid composition, on the protein predicted secondary structure, as predicted by SSpro, and on a boosted linear combination of HMM e-values obtained from 3,380 HMMs built from multiple sequence alignments of specific fragments - called contact fragments - of both capsid and tail sequences. This tool was designed in the context of a 2015 Explorer Program visit at University of California, Irvine

- Participants: Clovis Galiez, François Coste
- Partner: Pierre Baldi, University of California, Irvine
- Contact: Clovis Galiez

6.3. Shogen

KEYWORDS: Systems Biology - Bioinformatics - Genomics
FUNCTIONAL DESCRIPTION

This ASP-based software aims at identifying every segments of consecutive genes in a bacterial genome with a maximum number of genes that participates in a given metabolic pathway. Through this selection, the shogen tool deciphers putative sets of genes that (1) take an active part in metabolic pathways while being closely connected via metabolic networks and (2) are consecutive on each of the genomes involved. In practice, our approach connects genomic and metabolic knowledge by considering the genome organization and the biochemical reactions catalyzed by enzymes encoded by its genes. The underline parsimonious principle assumes that genes must be jointly regulated to activate a metabolic reaction cascade, and should be close enough in the genome organization. In 2015, the tool was simplified to handle standardized data formats, enabling its application to the modelling of a bacterial community [17]

- Participants: Philippe Bordron, Damien Eveillard, Alejandro Maass and Anne Siegel
- Partners: LINA - University of Chile
- Contact: Anne Siegel
- URL: http://aspforbiology.genouest.org/wiki.php/Software%20&%20Biological%20applications

6.4. Caspo

Cell ASP Optimizer
FUNCTIONAL DESCRIPTION

Cell ASP Optimizer (caspo) is a pipeline for automated reasoning on logical signaling networks. The main underlying issue is that inherent experimental noise is considered, so that many different logical networks can be compatible with a set of experimental observations. Caspo-learn performs an automated inference of logical networks from experimental data. It identifies admissible large-scale families of logic models without any a priori bias, thus saving a lot of efforts. Next, once a family a logical networks has been identified, caspo-design can suggest or design new experiments in order to reduce the uncertainty associated to this family. Finally, caspo-control computes 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. In 2015, the tool was extended to compute experimental design proposition [23], and to handle time-series datasets [31].

- Participants: Santiago Videla, Carito Guziolowski, Sven Thiele, Thomas Cokelaer, Torsten Schaub, Anne Siegel, Loic Paulevé and Max Ostrowski
- Partners: Ecole Centrale de Nantes - University of Potsdam - EMBL - LRI - Laboratoire de Recherche en Informatique
- Contact: Anne Siegel
- URL: http://bioasp.github.io/caspo/

6.5. 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 contains complementary tools with respect to their targeted sub-domain of bioinformatics.

6.5.1. 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 and solve incompleteness of large-scale metabolic networks. Since November 2015, a new version of Meneco has been available with Python 3, and a new functionality of topological producibility checking has been set up. [82] [60] [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. See section 6.3 for details. [55][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. In 2015, the tool was simplified to handle standardized data formats. [41] [web server].

ingranalysis [*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. The tool has evolved to the *iggy* tool recently [5] [21] [Python package] [web server].

6.5.2. 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. See Sec. 6.4 for details. The tool is included in the cellNopt package [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 [40][webserver][software and doc].

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

6.5.3. Sequence annotation

We develop tools for discovery and search of complex signatures within biological 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 [48]. In 2015, the efficiency of the tool was improved by slight evolutions of the underlying grammar. Possible fields of application are the detection of mutated binding sites or stem-loop identification (e.g. in CRISPR [2] [10]) [software].

Protomata learner Protomata software suite provides a grammatical inference framework for learning the specific signature of a functional protein family from unaligned sequences by partial and local multiple alignment and automata modeling. In 2015, motivated by the characterization of viral protein sequences during the internship of Maud Jusot [38], we have begun a refactoring of the parsing part of Protomata and we implemented a new mode returning the sum of the scores over all paths (Forward score), besides the classical score on best path (Viterbi score), to improve parsing's sensitivity on divergent but conserved families of sequences. [web server].

[1] http://www.cellnopt.org/
[2] http://crispi.genouest.org/

6.5.4. Integration of toolboxes and platforms in webservices

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

Plateform for data storage, expertise sharing and application inventory In collaboration with the GenOuest ressource center, the BII plateform (Bio Investigation Index) is a good way to enhance knowledge and expertise sharing, improve the visibility on the team's work in progress and record the History of the team's discoveries and main results. It enables experiment reproducibility, reporting on experiment process details, storing all scripts and softwares (in the corresponding versions) and linking all input files, results and not reproducible intermediate data. [web access].

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.

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].

7. New Results

7.1. Data integration

Participants: Jacques Nicolas, Charles Bettembourg, Jérémie Bourdon, Jeanne Got, Marie Chevallier, Guillaume Collet, Olivier Dameron, Damien Eveillard, Julie Laniau, Anne Siegel.

Extended notions of sign consistency to relate experimental data to signaling and regulatory network topologies. Interaction graphs provide a suitable representation of cellular networks with information flows. Methods based on sign consistency have been shown to be valuable tools to (i) predict qualitative responses, (ii) test the consistency of network topologies and experimental data, and (iii) apply repair operations to the network model suggesting missing or wrong interactions. We present a framework to unify different notions of sign consistency and propose a refined method for data discretization that considers uncertainties in experimental profiles. We furthermore introduce a new constraint to filter undesired model behaviors induced by positive feedback loops. Finally, we generalize the way predictions can be made by the sign consistency approach. This corresponds to an extension of our *Bioquali* software. [*Anne Siegel*] [21]

Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. Our software tool *shogen* was used to decipher functional roles within a consortium of five mining bacteria through the integration of genomic and metabolic knowledge at genome scale. We first reconstructed a global metabolic network. Next, using a parsimony assumption, we deciphered sets of genes, called Sets from Genome Segments (SGS), that (i) are close on their respective genomes, (ii) take an active part in metabolic pathways and (iii) whose associated metabolic reactions are also closely connected within metabolic networks. The use of SGS (shogen) pinpoints a functional compartmentalization among the investigated species and exhibits putative bacterial interactions necessary for promoting these pathways. [*Damien Eveillard, Anne Siegel*] [17]

Optimal Threshold Determination for Interpreting Semantic Similarity and Particularity: Application to the Comparison of Gene Sets and Metabolic Pathways Using GO and ChEBI. We developed a method for determining optimal semantic similarity and particularity thresholds in order to interprete the results of the comparison of ontology terms sets. We applied this method on the GO and ChEBI ontologies. Qualitative analysis using the thresholds on the PPAR multigene family yielded biologically-relevant patterns. [*Charles Bettembourg, Olivier Dameron*] [16]

AskOmics : Integration et interrogation de reseaux de regulation genomique et post-genomique. We present AskOmics, an integration and interrogation software using a RDF model and the SPARQL query language. The purpose of this work is to obtain quick answers to biological questions demanding currently hours of manual search in several spreadsheet results files. AskOmics allows biologists to integrate and interrogate their data by themselves without any knowledge about RDF and SPARQL required. [*Charles Bettembourg, Olivier Dameron*] [30]

7.2. Time-series and asymptotic dynamics

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

Identification of logical models for signaling pathways: towards a systems biology loop. Logical models of signaling pathways are a promising way of building effective in silico functional models of a cell. The automated learning of Boolean logic models describing signaling pathways can be achieved by training to phosphoproteomics data. This data is unavoidably subject to noise. As a result, the learning process leads to a family of feasible logical networks rather than a single model. This family is composed of logic models proposing different internal wirings for the system, implying that the logical predictions from this family may suffer a significant level of variability leading to uncertainty. In our work, combinatorial optimization methods based on recent logic programming paradigm allow to enumerate, and discriminate the family of logical models explaining data. Together, these approaches enable a robust understanding of the system response. The results are implemented in the *caspo* software [*Jacques Nicolas, Anne Siegel*] [22], [23]

Boolean Network Identification from Multiplex Time Series Data. The ASP-based learning algorithm developed in the team to train logical models of signaling networks focuses on the comparison of two time-points and assumes that the system has reached an early steady state. We have generalized such a learning procedure in order to discriminate Boolean networks according to their transient dynamics. To that goal, we exhibit a necessary condition that must be satisfied by a Boolean network dynamics to be consistent with a discretized time series trace. This approach was included in the ASP-based framework designed for the caspo software. We ended up with a global learning algorithm and compared it to learning approaches based on static data. [*Anne Siegel*] [31]

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

families of substitutions related to multi-dimensional continued fraction algorithms. [*Anne Siegel*] [32], [25], [14]

Multivariate Normal Approximation for the Stochastic Simulation Algorithm: Limit Theorem and Applications. We present a central limit theorem for the Gillespie stochastic trajectories when the living system has reached a steady-state, that is when the internal bio-molecules concentrations are assumed to be at equilibrium. It appears that the stochastic behavior in steady-state is entirely characterized by the stoichiometry matrix of the system and a single vector of reaction probabilities. We propose several applications of this result such as deriving multivariate confidence regions for the time course of the system and a constraints-based approach which extends the flux balance analysis framework to the stochastic case. [*Jérémie Bourdon, Vincent Picard, Anne Siegel*] [20], [12]

A Logic for Checking the Probabilistic Steady-State Properties of Reaction Networks. Designing probabilistic reaction models and determining their stochastic kinetic parameters are major issues in systems biology. In order to assist in the construction of reaction network models, we introduce a logic that allows one to express asymptotic properties about the steady-state stochastic dynamics of a reaction network. Basically, the formulas can express properties on expectancies, variances and co-variances. We demonstrate that deciding the satisfiability of a formula is NP-hard. [*Jérémie Bourdon, Vincent Picard, Anne Siegel*] [28], [12]

7.3. Sequence and structure annotation

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

Amplitude Spectrum Distance: measuring the global shape divergence of protein fragments. We introduce here the Amplitude Spectrum Distance (ASD), a novel way of comparing protein fragments based on the discrete Fourier transform of their C_α distance matrix. Defined as the distance between their amplitude spectra, ASD can be computed efficiently and provides a parameter-free measure of the global shape dissimilarity of two fragments. ASD inherits from nice theoretical properties, making it tolerant to shifts, insertions, deletions, circular permutations or sequence reversals while satisfying the triangle inequality. The practical interest of ASD with respect to RMSD, RMSD_d, BC and TM scores is illustrated through zinc finger retrieval experiments and concrete structure examples. The benefits of ASD are also illustrated by two additional clustering experiments: domain linkers fragments and complementarity-determining regions of antibodies. [*Clovis Galiez, François Coste*] [19]

Structural conservation of remote homologues: better and further in contact fragments. We address a basic question on sequence-structure relationships in proteins: does a protein sequence depict a structure with a uniform faithfulness all along the sequence ? We investigate this question by defining contact fragments. This study suggests that sequence homologs of CF are significantly more faithful to structure than randomly chosen fragments, so that CF carry a strong sequence-structure relationship, allowing them to be used as accurate building blocks for structure prediction. [*Clovis Galiez, François Coste*] [26]

VIRALpro: a tool to identify viral capsid and tail sequences. Not only sequence data continues to outpace annotation information, but the problem is further exacerbated when organisms are underrepresented in the annotation databases. This is the case with non human-pathogenic viruses which occur frequently in metagenomic projects. Thus there is a need for tools capable of detecting and classifying viral sequences. We describe VIRALpro a new effective tool for identifying capsid and tail protein sequences, which are the cornerstones toward viral sequence annotation and viral genome classification. [*Clovis Galiez, François Coste*] [18]

Finding Optimal Discretization Orders for Molecular Distance Geometry. The Molecular Distance Geometry Problem (MDGP) is the problem of finding the possible conformations of a molecule by exploiting available information about distances between some atom pairs. Under minimal assumptions the MDGP can be discretized so that the search domain of the problem becomes a tree that can be explored by using an interval Branch & Prune (iBP) algorithm. In this context, the discretization assumptions are strongly dependent on the atomic ordering, which can also impact the computational cost of the iBP algorithm. In this work, we propose

a new partial discretization order for protein backbones. This new atomic order optimizes a set of objectives that aim at improving the iBP performances. The optimization of the objectives is performed by Answer Set Programming (ASP), which allows to express the problem by a set of logical constraints. The comparison with previously proposed orders for protein backbones shows that this new discretization order makes iBP perform more efficiently. [*Jacques Nicolas*] [34]

From formal concepts to analogical complexes. Reasoning by analogy is an important component of common sense reasoning whose formalization has undergone recent improvements with the logical and algebraic study of the analogical proportion. The starting point of this study considers analogical proportions on a formal context. We introduce analogical complexes, a companion of formal concepts formed by using analogy between four subsets of objects in place of the initial binary relation. They represent subsets of objects and attributes that share a maximal analogical relation. We show that the set of all complexes can be structured in an analogical complex lattice and give explicit formulae for the computation of their infimum and supremum. [*Laurent Miclet, Jacques Nicolas*] [27]

Comparison of the targets obtained by a scoring matrix and by a regular expression. Application to the search for LXR binding sites. In bioinformatics, it is a common task to search for new instances of a pattern built from a set of reference sequences. For the simplest and most frequent cases, patterns are represented in two ways : regular expression or scoring matrix. Since both representations seem to be used indifferently in pratice, one may wonder if they have any impact on the result. This study compares hits obtained with scoring matrices or by regular expressions allowing up to two substitutions. It shows that, in our LXR study, sequences found by a scoring matrix are closer to the targeted hits than sequences found by a regular expression. [*Aymeric Antoine-Lorquin, Jacques Nicolas, Catherine Belleannée*] [29]

Finding and Characterizing Repeats in Plant Genomes. Plant genomes contain a particularly high proportion of repeated structures of various types. This chapter proposes a guided tour of available softwares that can help biologists to look for these repeats and check some hypothetical models intended to characterize their structures. Since transposable elements are a major source of repeats in plants, we have provided a whole section on this topic as well as a selection of the main existing softwares. In order to better understand how they work and how repeats may be efficiently found in genomes, the rest of the chapter is devoted to the foundations of the search for repeats and more complex patterns. We first introduce the key concepts that are useful for understanding the current state of the art in playing with words, applied to genomic sequences. In fact, biologists need to represent more complex entities where a repeat family is built on more abstract structures, including direct or inverted small repeats, motifs, composition constraints as well as ordering and distance constraints between these elementary blocks. The last section introduces concepts and practical tools that can be used to reach this syntactic level in biological sequence analysis. [*Jacques Nicolas*] [35]

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Regional partnership with computer science laboratories in Nantes

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

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.

8.1.2. Regional partnership in Marine Biology

Participants: Catherine Belleannée, Jérémie Bourdon, Guillaume Collet, Jean Coquet, François Coste, Damien Eveillard, Olivier Dameron, Clémence Frioux, Clovis Galiez, Jeanne Got, Yann Guitton, Julie Laniau, Jacques Nicolas, Vincent Picard, Camille Trottier, Anne Siegel.

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

8.1.3. 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, Nathalie Théret, Aurélie Evrard.

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-doctoral student and the co-supervision of several PhD students. The Ph-D thesis of V. Wucher was supported by collaborations with the IGEPP laboratory. The post-doc of Charles Bettembourg strengthens these collaborations. This collaboration is also reinforced by collaboration within ANR contracts (MirNadapt, FatInteger). Lately, Aurélie Evrard joined the team at mid-part of her time in collaboration with Agrocampus Ouest and INRA to apply the semantic web to technologies developed within the mirnAdapt framework to new agriculture applications (Brassicae).

We also have a strong and long term collaboration in the bio-medical domain, namely with the IRSET laboratory at Univ. Rennes 1/Irset, acted by the co-supervised Ph-D theses of V. Delannée (Metagenotox project, funded by Anses) and J. Coquet. This partnership was reinforced in the former years by the ANR contract Biotempo ended at the end of 2014. In 2015, the project of combining semantic web technologies and bi-clustering classification based on formal concept analysis was applied to systems biology within the PEPS CONFOCAL project. This scientific project will be pushed forward in the recent TGFSYSBio project funded by Plan Cancer on the modelling of the microenvironment of TGFbeta signaling network.

8.2. National Initiatives

8.2.1. Long-term contracts

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

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

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.

8.2.1.2. ANR Idealg

Participants: Jérémie Bourdon, Marie Chevallier, Guillaume Collet, François Coste, Damien Eveillard, Clémence Frioux, Clovis Galiez, Jeanne Got, Yann Guitton, Jacques Nicolas, Anne Siegel.

IDEALG is one of the five laureates from the national call 2010 for Biotechnology and Bioresource and will run until 2020. It gathers 18 different partners from the academic field (CNRS, IFREMER, UEB, UBO, UBS, ENSCR, University of Nantes, INRA, AgroCampus), the industrial field (C-WEED, Bezhin Rosko, Aleor, France Haliotis, DuPont) as well as a technical center specialized in seaweeds (CEVA) in order to foster biotechnology applications within the seaweed field. It is organized in ten workpackages. We are participating 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 [More details].

8.2.2. Proof-of-concept on dedicated applications

8.2.2.1. 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.

8.2.2.2. ANR Mirnadapt

Participants: Jacques Nicolas, Anne Siegel, Olivier Dameron, 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 [92]. 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.

8.2.2.3. ANR Samosa

Participants: Anne Siegel, Jeanne Got, Damien Eveillard.

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

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

8.2.3. Programs funded by research institutions

8.2.3.1. INSERM TGFSYSBIO

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

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

8.2.3.2. ADT Complex-biomarkers and ADT Proof of concept

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

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

8.2.3.3. ANSES Mecagenotox

Participants: Victorien Delannée, Anne Siegel, Nathalie Théret.

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

8.2.3.4. PEPS VAG

Participants: François Coste, Clovis Galiez, Jacques Nicolas.

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. More specifically,we develop new methods based on both sequential and structural information of proteins to improve the detection of viral sequences in marine metagenomes. This will make possible 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.

8.2.3.5. PEPS CONFOCAL

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

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

8.3. European Initiatives

8.3.1. 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 modeling and study of biological networks.

8.4. International Initiatives

8.4.1. Inria International Labs

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

Inria Chile

Associate Team involved in the International Lab:

8.4.1.1. BIOINTEGRATIVECHILE

Title: Integrative Biology in Extreme Environments

International Partner (Institution - Laboratory - Researcher):

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

Start year: 2014

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

The project is in the area of bioinformatics, with a special focus on bacteria living in extreme environments, more precisely on microorganisms involved in bio-remediation or bio-production processes. We are particularly interested in bioprocesses such as copper extraction, salmon lethality, metal-resistance, all having an economical interest in Chile. Since the last decade, huge databases of microbial genomic sequences, together with multi-scale and large-scale cellular observations (genomics, transcriptomics, proteomics, metabolomics) have been produced. Each one can be considered as a different scale of a biological process, either in time or space. But ultimately they are related through networks of biological interactions that control the behavior of the system. The reconstruction, analysis and modeling of such networks using all levels of information are

biologically, mathematically and computationally challenging. Applied on microorganisms living in extreme environments, this question is even more challenging since relatively few knowledge is publicly available on the species, requiring to develop methods which are robust to uncertainty. We are developing methods to integrate and manage heterogeneous omics and uncertain data, this in the purpose of extracting suitable biomarkers from this multi-scale information. This question will be addressed by coupling probabilistic and static dynamical systems methods with recent and efficient paradigms of constraint programming (Answer Set Programming).

8.4.2. Inria Associate Teams

8.4.2.1. 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.

8.5. International Research Visitors

8.5.1. Visits of International Scientists

- **Chile**. Centro de Modelimiento Matematico, Santiago [A. Maass, N. Loirà, M. Latorre]
- **Germany**. Frei Universitat Berlin [A. Bockmayr, H. Siebert]
- **Niger**. University of Maradi [O. Abdou-Arbi]
- **Turkey**. University of Istambul [A. Aravena]

8.5.2. Visits to International Teams

8.5.2.1. Explorer program

Galiez Clovis

Date: Mar 2015 - May 2015

Institution: University of California, Irvine (United States)

8.5.2.2. Short visits

- **Chile**. Centro de Modelimiento Matematico, Santiago de Chile [J. Bourdon, M. Chevallier, C. Frioux, A. Siegel]
- **Chile**. Centro de Modelimiento Matematico, Santiago de Chile [M. Chevallier]
- **Germany**. Frei Berlin University [A. Siegel]

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events selection

9.1.1.1. Member of the conference program committees

- CMSB (2015): Computational Methods on Systems Biology [A. Siegel]
- BBCC (2015): Bioinformatica e Biologia Computazionale in Campania [O. Dameron]
- SIIM (2015): Symposium sur l'Ingénierie des Informations Médicales [O. Dameron]

9.1.2. Journal

9.1.2.1. Member of the editorial boards

- Academic editor: Plos One [J. Bourdon]

9.1.2.2. Reviewer - Reviewing activities

- Journal of Mathematical Biology. Theorie des Sciences Informatiques. [A. Siegel]
- Theoretical Computer Science, Fundamenta Informaticae. [F. Coste]
- 2014 O. Dameron: Bioinformatics, Cancer informatics, Journal of Biomedical Informatics, Journal of Biomedical Semantics.
- 2015 O. Dameron: Journal of Biomedical Semantics (x3).

9.1.3. Invited talks

- A. Antoine-Lorquin *Orthocis : Une base de données pour l'étude des facteurs de transcriptions. Identification in silico de gènes ciblés par un facteur de transcription donné*. UMR MIA. AgroParis-Tech (Sep. 2015)
- M. Chevallier *Systems biology : uses of the Mobyle and Galaxy plateforms*. Bio-informatics day. e-Biogenouest and GRIOTE. (Avr. 2015).
- F. Coste *Learning efficiently (local) substitutable context-free languages from text*. XRCE seminar, Grenoble. (Jul. 2015)
- F. Coste *Learning protein languages*. Machine learning thematic trimester. CIMI, Toulouse. (Dec. 2015)
- O. Dameron *Knowledge-based selection of candidate metabolic networks*. CRC (équipe A Burgun). (Feb. 2015)
- C. Galiez *Sequence-structure relationship in protein sequences and applications to sequence annotation.* LBCQ (équipe A. Carbone), Paris. (Juil. 2015)
- C. Galiez *Apprentissage de prédictions structurales depuis la séquence pour l'annotation fonctionnelle.* LIRMM (équipe O. Gascuel), Montpellier. (Oct. 2015)
- C. Galiez *Structural fragments : comparison, predictability from sequence and application to identification of viral proteins.* MPI (équipe J. Söding), par Internet, Göttingen. (Dec. 2015)
- Y. Guitton *Metabolic profiling and control hypothesis through kinetic accumulation or elimination of secondary metabolites associated with phycotoxins of filter-feeding bivalves* RFMF 2015 (9. édition des Journées Scientifiques du Réseau Francophone de Métabolomique et Fluxomique). Lille. June 2015.
- Y. Guitton *L'annotation automatisée des analyses métabolomiques en LC-MS : un challenge relevé par la plate-forme Corsaire*. Gen2Bio. Mars 2015.
- J. Laniau *Combinatorial optimization methods to complete and analyse a metabolic network*. Journée BIOSS. Nantes. Septembre 2015
- J. Laniau *Combinatorial optimization methods to complete and analyse a metabolic network*. Journée MOABI. Paris. Novembre 2015.
- V. Picard *Asymptotic Analysis of Gillespie Algorithm under Steady-State Assumption*. Advanced Lecture Course on Computational Systems Biology. Aussois (Apr. 2015)
- V. Picard *Analyse stationnaire des réseaux de réactions : systèmes de contraintes en modélisation stochastique*. Paris. Journées du groupe de travail bioss (Nov. 2015).

- A. Siegel *Numeration, redundancy graphs and topological properties of fractals*. Department of Mathematics. Université Paris Sud (Jan. 2015)
- A. Siegel *Topological properties of generalized Rauzy fractals: which novel issues?*. FAN numeration meeting. Admont, Austria. (Jun. 2015).
- A. Siegel *Decidability problems for self-induced systems generated by a substitution?*. Conference MCU'2015. North Cyprus. (Sept. 2015).
- A. Siegel *Confronting knowledge networks on signaling networks with phosphoproteomics datasets using combinatorial optimization approaches*. Departement of computer sciences. Frei Berlin University. (Oct. 2015).
- A. Siegel *Data science and systems biology*. Data science meeting. IRISA, Rennes (Nov. 2015).

9.1.4. Leadership within the scientific community

- Member of the steering committee of the International Conference on Grammatical Inference [F. Coste].
- The team was involved in the foundation of a national working group on the symbolic study of dynamical systems named bioss [web access]. The group gathers more 100 scientists, from computer science to biology. Three meetings were organized this year. The group is supported by two French National Research Networks: bioinformatics (GDR BIM : bioinformatique moléculaire) and informatics-mathematics (GDR IM : Informatique Mathématique). [A. Siegel]

9.1.5. Scientific expertise

- Scientific Advisory Board of GDR BIM " Molecular Bioinformatics"[J. Nicolas].
- Inria National evaluation board [A. Siegel]
- Member (nominated) CNU section 65 [O. Dameron]
- Member of the Operational Legal and Ethical Risk Assessment Committee (COERLE) at Inria [J. Nicolas].
- Recruitment committees: Professor (CRISTAL, Lille) [A. Siegel], Professor (IRHS, Angers) [A. Siegel], Inria junior researcher (Nice) [A. Siegel].
- Member of the IRISA laboratory council [F. Coste].
- Member of the Inria Rennes center council [A. Siegel].
- Scientific Advisory Board of Biogenouest [J. Bourdon, A. Siegel].
- Member of SCAS (Service Commun d'Action Sociale) of Univ. Rennes 1 [C. Belleannée].
- Member of CUMIR (Commission des Utilisateurs des Moyens Informatiques, Inria Rennes) [F. Coste].

9.2. Teaching - Supervision - Juries

9.2.1. 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.

O. Dameron shares the coordination of the "Bioinformatique et génomique" Master degree, Univ. Rennes1, and of the "Méthodes et traitements de l'information biomédicale et hospitalière"

O. Dameron shares the coordination of the course "Bioinformatique expérimentale", Master1 in computer science, Univ. Rennes1 + ENS. He is the coordinator of the courses "Bases de mathématiques et probablité" and "Méthodes en informatique" in Master1 in public health (Univ. Rennes 1), "Représentation des connaissances biomédicales" in Master2 in public health (Univ. Rennes 1), "Principes de programmation et d'algorithmique" and "Gestion de projets informatiques" in Master1 in bioinformatics (Univ. Rennes 1), "Standardisation des connaissances et bio-ontologies" in Master2 in bioinformatics (Univ. Rennes1), as well as of the course "e-Santé et réseaux hospitaliers" in the last year if engineering school ESIR (Univ. Rennes 1).

9.2.2. Teaching

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

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

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

Licence: V. Delannée, Optique, 4h, PACES, Univ. Rennes 1, France.

Licence: V. Delannée, Biostatistiques, 24h, PACES, Univ. Rennes 1, France.

Licence: J. Coquet, Algorithmique et Programmation Fonctionnelle, 20h, L1 informatique, Rennes1, France.

Licence: J. Coquet, Bureautique, 10h, L1 informatique, Rennes1, France.

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

Licence: V. Picard, Probability theory, 24h, L3, ENS Rennes, France.

Master: J. Coquet, Principes de programmation et d'algorithmique, 32h, M1 BioInformatique et génomique, Rennes1, France.

Master: A. Antoine-Lorquin, Principes de programmation et d'algorithmique, 32h, M1 BioInformatique et génomique, Univ. Rennes 1, France.

Master: A. Antoine-Lorquin, Bases de mathématiques, probabilités et statistiques, 32h, M1 Master STS - Mention Santé publique, Univ. Rennes 1, France.

Master: C. Belleannée, algorithmique du texte et bioinformatique, M1 informatique, Univ. Rennes1, France

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

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

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

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

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

Master: O. Dameron, Standardisation des connaissances et bio-ontologies, 22h, M2 bioinformatique et génomique, 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, 16h, M1, ENS Rennes/Univ. Rennes 1, France.

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

Engineer: O. Dameron, e-Santé et réseaux hospitaliers, ESIR, Rennes.

Engineer: O. Dameron, Bio-ontologies et Web Sémantique, ENSTBr, Brest.

9.2.3. Supervision

PhD : Clovis Galiez, *Structural fragments : comparison, predictability from sequence and application to identification of viral proteins*, 8 Dec. 2015, supervised by F. Coste and J. Nicolas. [11]

PhD: Vincent Picard, *Analyse dynamique d'algorithmes et dynamique symbolique pour l'étude de modèles semi-quantitatifs en biologie des systèmes*,16 Dec. 2015, supervised by A. Siegel and J. Bourdon. [12]

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 : 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.

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

9.2.4. Juries

- *Member of Ph-D thesis jury*. T. Nguyen, LABRI, Bordeaux [A. Siegel, rapporteure]. N. Mobillia, IMAG, Grenoble [A. Siegel, présidente].
- *Member of medical thesis jury*. V. Margot, Rennes [O. Dameron, examinateur].

9.2.5. Internships

- Internship, from February until July 2015. Supervised by A. Siegel. Student: Clémence Frioux. Subject: Iterative reconstruction of functional metabolic networks for non-model organisms
- Internship, from April until July 2015. Supervised by O. Dameron and J. Coquet. Student: Pierre Vignet. Subject: Evaluation of a knowledge-based selection strategy for candidate metabolic pathways
- Internship, from January until July 2015. Supervised by F. Coste. Student: Maud Jusot. Subject: Syntactic modelling of viral genomes
- Internship, from April until July 2015. Supervised by A. Siegel. Student: Meziane Aite. Subject: Tools for the reconstruction of the metabolic network of T. Lutea
- Internship, from April until July 2015. Supervised by J. Nicolas. Student: Lucas Bourneuf. Subject: Model reduction with power graph algorithms
- Internship, from April until Aug 2015. Supervised by C. Belleannée. Student: David Picard-Druet. Subject: Coupling pattern discovery and pattern matching to design a composite signature for the Yeast polyadenylation site

9.3. Popularization

- *Organization of Sciences en Cour[t]s*. Popularization Festival where PhD students explain their thesis via short films. [*J. Coquet, V. Delannée, A. Antoine-Lorquin, C. Bettembourg*] [more info].
- *Production of Sciences en Cour[t]s film*. "Une rencontre percutante": A short movie about metabolic reconstruction.[*J. Coquet, V. Delannée, A. Antoine-Lorquin*] [more info].

- *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, file versionning with GIT. [*G. Collet, O. Dameron*]. [more info].
- *Animation of Bioinformatics Atelier at Data Science Symposium (IRISA's 40th anniversary)* "De la bioinformatique aux tiques". [*F. Coste*]
- *Participation to the Data Science Symposium (IRISA's 40th anniversary)* "Science data ecosystem". [*A. Siegel, O. Dameron*]

10. Bibliography

Major publications by the team in recent years

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

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

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

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

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

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

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

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

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

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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] C. GALIEZ. *Structural fragments : comparison, predictability from the sequence and application to the identification of viral structural proteins*, Université de Rennes 1, December 2015, https://hal.archives-ouvertes.fr/tel-01243132

[12] V. PICARD. *Reaction networks : from probabilistic analysis to refutation*, Université de Rennes 1, December 2015, https://hal.inria.fr/tel-01246180

Articles in International Peer-Reviewed Journals

[13] V. ACUÑA, A. ARAVENA, C. GUZIOLOWSKI, D. EVEILLARD, A. SIEGEL, A. MAASS. *Deciphering transcriptional regulations coordinating the response to environmental changes*, in "BMC Bioinformatics", January 2015, vol. 17, n^{o} 1, 35 p. , https://hal.archives-ouvertes.fr/hal-01260866

[14] V. BERTHÉ, J. BOURDON, T. JOLIVET, A. SIEGEL. *A combinatorial approach to products of Pisot substitutions*, in "Ergodic Theory and Dynamical Systems", 2015, 38 p. [*DOI :* 10.1017/ETDS.2014.141], https://hal.inria.fr/hal-01196326

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

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

[17] P. BORDRON, M. LATORRE, M.-P. CORTES, M. GONZALES, S. THIELE, A. SIEGEL, A. MAASS, D. EVEILLARD. *Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach*, in "MicrobiologyOpen", 2015 [*DOI :* 10.1002/MBO3.315], https://hal.inria.fr/hal-01246173

[18] C. GALIEZ, M. CHRISTOPHE, F. COSTE, P. BALDI. *VIRALpro: a tool to identify viral capsid and tail sequences*, in "Bioinformatics", December 2015, https://hal.archives-ouvertes.fr/hal-01242251

[19] C. GALIEZ, F. COSTE. *Amplitude spectrum distance: measuring the global shape divergence of protein fragments*, in "BMC Bioinformatics", August 2015, vol. 16, n^{o} 1, 16 p. [*DOI :* 10.1186/S12859-015-0693-Y], https://hal.inria.fr/hal-01214482

[20] V. PICARD, A. SIEGEL, J. BOURDON. *Multivariate Normal Approximation for the Stochastic Simulation Algorithm: Limit Theorem and Applications*, in "Electronic Notes in Theoretical Computer Science", 2015, vol. 316C, pp. 67-82 [*DOI :* 10.1016/J.ENTCS.2015.06.011], https://hal.inria.fr/hal-01196533

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

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

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

Invited Conferences

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

[25] T. JOLIVET, A. SIEGEL. *Decidability Problems for Self-induced Systems Generated by a Substitution*, in "MCU 2015 : Machines, Computation and Universality (7th conference)", Famagusta, Cyprus, Springer, 2015, vol. Lecture Notes in Computer Science 9288 [*DOI :* 10.1007/978-3-319-23111-2], https://hal.inria.fr/hal-01196152

International Conferences with Proceedings

[26] C. GALIEZ, F. COSTE. *Structural conservation of remote homologues: better and further in contact fragments*, in "3DSIG: Structural Bioinformatics and Computational Biophysics", Dublin, France, July 2015, 1 p. , https://hal.inria.fr/hal-01214506

[27] L. MICLET, J. NICOLAS. *From formal concepts to analogical complexes*, in "CLA 2015", Clermont-Ferrand, France, LIMOS, CNRS et Université Blaise Pascal, October 2015, 12 p. , https://hal.inria.fr/hal-01198943

[28] V. PICARD, A. SIEGEL, J. BOURDON. *A Logic for Checking the Probabilistic Steady-State Properties of Reaction Networks*, in "IJCAI workshop BAI: Bioinformatics and Artificial Intelligence", Buenos Aeres, Argentina, CEUR-WS, 2015, https://hal.inria.fr/hal-01196598

Conferences without Proceedings

[29] A. ANTOINE-LORQUIN, S. LAGARRIGUE, F. LECERF, J. NICOLAS, C. BELLEANNÉE. *Comparison of the targets obtained by a scoring matrix and by a regular expression. Application to the search for LXR binding sites*, in "JOBIM 2015- 16e Journées Ouvertes en Biologie, Informatique et Mathématiques", Clermont-Ferrand, France, July 2015, https://hal.inria.fr/hal-01197050

[30] C. BETTEMBOURG, O. DAMERON, A. BRETAUDEAU, F. LEGEAI. *AskOmics : Intégration et interrogation de réseaux de régulation génomique et post-génomique*, in "IN OVIVE (INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement)", Rennes, France, June 2015, 7 p. , https://hal.inria.fr/hal-01184903

[31] M. OSTROWSKI, L. PAULEVÉ, T. SCHAUB, A. SIEGEL, C. GUZIOLOWSKI. *Boolean Network Identification from Multiplex Time Series Data*, in "CMSB 2015 - 13th conference on Computational Methods for Systems Biology", Nantes, France, O. ROUX, J. BOURDON (editors), Lecture Notes in Computer Science, Springer International Publishing, September 2015, vol. 9308, pp. 170-181 [*DOI :* 10.1007/978-3-319-23401-4_15], https://hal.archives-ouvertes.fr/hal-01164751

Scientific Books (or Scientific Book chapters)

[32] S. AKIYAMA, M. BARGE, V. BERTHÉ, J.-Y. LEE, A. SIEGEL. *On the Pisot Substitution Conjecture*, in "Mathematics of Aperiodic Order", Progress in mathematics, Springer, 2015, vol. 309, pp. 33-72 [*DOI :* 10.1007/978-3-0348-0903-0_2], https://hal.inria.fr/hal-01196318

[33] F. COSTE. *Learning the Language of Biological Sequences*, in "Topics in Grammatical Inference", J. HEINZ, J. M. SEMPERE (editors), Springer-Verlag Berlin Heidelberg, 2016, https://hal.inria.fr/hal-01244770

[34] D. S. GONÇALVES, J. NICOLAS, A. MUCHERINO, C. LAVOR. *Finding Optimal Discretization Orders for Molecular Distance Geometry by Answer Set Programming*, in "Studies in Computational Intelligence", S. FIDANOVA (editor), Recent Advances in Computational Optimization, Springer, July 2015, vol. 610, pp. 1-15, https://hal.inria.fr/hal-01196714

[35] J. NICOLAS, P. PETERLONGO, S. TEMPEL. *Finding and Characterizing Repeats in Plant Genomes*, in "Plant Bioinformatics: Methods and Protocols", D. EDWARDS (editor), Methods in Molecular Biology, Humana Press - Springer Science+Business Media, November 2015, n^{o} 1374, 365 p. [*DOI :* 10.1007/978-1-4939-3167-5_17], https://hal.inria.fr/hal-01228488

Other Publications

[36] J. COQUET, G. ANDRIEUX, J. NICOLAS, O. DAMERON, N. THÉRET. *Topological and semantic Web based method for analyzing TGF-β signaling pathways*, December 2015, JOBIM 2015, Poster, https://hal.archives-ouvertes.fr/hal-01242893

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

[38] M. JUSOT. *Caractérisation en séquence et en structure des protéines virales*, UP7 - Université Paris Diderot - Paris 7, June 2015, https://hal.inria.fr/hal-01246372

[39] M. PETERA, G. LE CORGUILLE, M. LANDI, M. MONSOOR, M. TREMBLAY FRANCO, C. DUPERIER, J.-F. MARTIN, D. JACOB, Y. GUITTON, M. LEFEBVRE, E. PUJOS-GUILLOT, F. GIACOMONI, E. THÉVENOT, C. CARON. *Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics*, July 2015, JOBIM 2015 (16. édition des Journées Ouvertes en Biologie, Informatique et Mathématiques), Poster - Paper reference: Giacomoni et al. (2014) Workflow4Metabolomics: A collaborative research infrastructure for com-putational metabolomics. Bioinformatics http://dx.doi.org/10.1093/bioinformatics/btu813, https://hal.archives-ouvertes.fr/hal-01214152

References in notes

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

[41] 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

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

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

[44] J. B. L. BARD, S. Y. RHEE. *Ontologies in biology: design, applications and future challenges*, in "Nature reviews. Genetics", 2004, vol. 5, n° 3, pp. 213–222

[45] 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

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

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

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

[49] 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

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

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

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

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

[54] O. BODENREIDER, R. STEVENS. *Bio-ontologies: current trends and future directions*, in "Briefings in Bioinformatics", 2006, vol. 7, n° 3, pp. 256–274

[55] 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

[56] N. CANNATA, E. MERELLI, R. B. ALTMAN. *Time to Organize the Bioinformatics Resourceome*, in "PLoS Computational Biology", 2005, vol. 1, n° 7, pp. 0531–0533

[57] N. CANNATA, M. SCHRÖDER, R. MARANGONI, P. ROMANO. *A Semantic Web for bioinformatics: goals, tools, systems, applications*, in "BMC bioinformatics", 2008, vol. 9 Suppl 4, S1 p.

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

[59] J. J. CIMINO, X. ZHU. *The practical impact of ontologies on biomedical informatics*, in "Methods of information in medicine", 2006

[60] 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

[61] 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

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

[63] 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

[64] S. M. DITTAMI, T. BARBEYRON, C. BOYEN, J. CAMBEFORT, G. COLLET, L. DELAGE, A. GOBET, A. GROISILLIER, C. LEBLANC, G. MICHEL, D. SCORNET, A. SIEGEL, J. E. TAPIA, 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

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

[66] 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

[67] 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

[68] D. P. HILL, N. ADAMS, M. BADA, C. BATCHELOR, T. Z. BERARDINI, H. DIETZE, H. J. DRABKIN, M. ENNIS, R. E. FOULGER, M. A. HARRIS, J. HASTINGS, N. S. KALE, P. DE MATOS, C. J. MUNGALL, G. OWEN, P. RONCAGLIA, C. STEINBECK, S. TURNER, J. LOMAX. *Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology*, in "BMC genomics", 2013, vol. 14, 513 p.

[69] 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

[70] 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

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

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

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

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

[75] 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

[76] L. J. G. POST, M. ROOS, M. S. MARSHALL, R. VAN DRIEL, T. M. BREIT. *A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data*, in "Bioinformatics (Oxford, England)", 2007, vol. 23, n^o 22, pp. 3080–3087

[77] 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

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

[79] 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

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

[81] A. RUTTENBERG, T. CLARK, W. BUG, M. SAMWALD, O. BODENREIDER, H. CHEN, D. DOHERTY, K. FORSBERG, Y. GAO, V. KASHYAP, J. KINOSHITA, J. LUCIANO, M. SCOTT MARSHALL, C. OGBUJI, J. REES, S. STEPHENS, G. T. WONG, E. WU, D. ZACCAGNINI, T. HONGSERMEIER, E. NEUMANN, I. HERMAN, K.-H. CHEUNG. *Advancing translational research with the Semantic Web*, in "BMC Bioinformatics", 2007, vol. 8, n^o 3

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

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

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

[85] 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

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

[87] R. STEVENS, C. A. GOBLE, S. BECHHOFER. *Ontology-based Knowledge Representation for Bioinformatics*, in "Briefings in bioinformatics", 2000, vol. 1, n^o 4, pp. 398–416

[88] 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

[89] 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

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

[91] 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

[92] 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