IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Ecole Polytechnique

Activity Report 2014

Project-Team AMIB

Algorithms and Models for Integrative Biology

IN COLLABORATION WITH: Laboratoire d'informatique de l'école polytechnique (LIX), Laboratoire de recherche en informatique (LRI)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Computational Biology

Table of contents

Project-Team AMIB

Keywords: Computational Structural Biology, Annotation, Systems Biology, Machine Learning, Algorithms

Creation of the Team: 2009 May 01*, updated into Project-Team:* 2011 January 01.

1. Members

Research Scientists

Mireille Régnier [Team leader, Inria, Senior Researcher, HdR]
Julie Bernauer [Inria, Researcher]
Loic Paulevé [CNRS, Researcher]
Yann Ponty [CNRS, Researcher]

Faculty Members

Patrick Amar [Univ. Paris Sud, Associate Professor, HdR]
Philippe Chassignet [Ecole Polytechnique, Associate Professor]
Alain Denise [Univ. Paris Sud, Professor, HdR]
Sarah Cohen-Boulakia [Univ. Paris Sud, Associate Professor]
Sabine Peres [IUT Orsay, Associate Professor]
Christine Froidevaux [Univ. Paris Sud, Professor, HdR]
Jean-Marc Steyaert [Ecole Polytechnique, HdR]

PhD Students

Erwan Bigan [Ecole Polytechnique]
Mélanie Boudard [Univ. Versailles and Univ. Paris-Sud]
Bryan Brancotte [Univ. Paris-Sud]
Adrien Guilhot-Gaudeffroy [Univ. Paris XI, until Sep 2014]
Alice Heliou [Ecole Polytechnique, from Jul 2014]
Amélie Heliou [Ecole Polytechnique, from Feb 2014]
Daria Iakovishina [Inria]
Cécile Pereira [Univ. Paris-Sud]
Adrien Rougny [Univ. Paris-Sud]
Vincent Le Gallic [Univ. Paris-Sud]
Antoine Soulé [Ecole Polytechnique]

Post-Doctoral Fellows

Olga Berillo [PostDoc, until Feb 2014]
Evgeniia Furletova [PostDoc IMPB, until Nov 2014]
Rasmus Fonseca [PostDoc,DIKU]

Visiting Scientist

Jan Holub [Professor, CTU FIT,from Sep 2014 until Oct 2014]

Administrative Assistant

Évelyne Rayssac [Ecole Polytechnique]

2. Overall Objectives

2.1. Overall Objectives

Our project addresses a central question in bioninformatics, namely the molecular levels of organization in the cells. The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. Therefore, folding and docking are still major issues in modern structural biology and we currently concentrate our efforts on structure, interactions, evolution and annotation and aim at a contribution to protein engineering and RNA design. With the recent development of molecular systems biology aiming to integrate different levels of information, protein and nucleic acid assemblies' studies should provide a better understanding on the molecular processes and machinery occurring in the cell and our research extends to several related issus in systems biology.

On the one hand, we study and develop methodological approaches for dealing with macromolecular structures and annotation: the challenge is to develop abstract models that are computationally tractable and biologically relevant. Our approach puts a strong emphasis on the modeling of biological objects using classic formalisms in computer science (languages, trees, graphs...), occasionally decorated and/or weighted to capture features of interest. To that purpose, we rely on the wide array of skills present in our team in the fields of combinatorics, formal languages and discrete mathematics. The resulting models are usually designed to be amenable to a probabilistic analysis, which can be used to assess the relevance of models, or test general hypotheses.

On the other hand, once suitable models are established we apply these computational approaches to several particular problems arising in fundamental molecular biology. One typically aims at designing new specialized algorithms and methods to efficiently compute properties of real biological objects. Tools of choice include exact optimization, relying heavily on dynamic programming, simulations, machine learning and discrete mathematics. As a whole, a common toolkit of computational methods is developed within the group. The trade-off between the biological accuracy of the model and the computational tractability or efficiency is to be addressed in a closed partnership with experimental biology groups. One outcome is to provide software or platform elements to predict either structures or structural and functional annotation. As members of the Inria community, we are part of the ADT BIOSCIENCES led by J. Nicolas whose goal is to develop a global INRIA Bioinformatics web portal.

3. Research Program

3.1. RNA

At the secondary structure level, we contributed novel generic techniques applicable to dynamic programming and statistical sampling, and applied them to design novel efficient algorithms for probing the conformational space. Another originality of our approach is that we cover a wide range of scales for RNA structure representation. For each scale (atomic, sequence, secondary and tertiary structure...) cutting-edge algorithmic strategies and accurate and efficient tools have been developed or are under development. This offers a new view on the complexity of RNA structure and function that will certainly provide valuable insights for biological studies.

3D modeling was supported by the Digiteo project JAPARIN-3D. Statistical potentials were supported by CARNAGE and ITSNAP.

3.1.1. Dynamic programming and complexity

Participants: Alain Denise, Yann Ponty, Antoine Soulé.

Common activity with J. Waldispühl (McGill).

Ever since the seminal work of Zuker and Stiegler, the field of RNA bioinformatics has been characterized by a strong emphasis on the secondary structure. This discrete abstraction of the 3D conformation of RNA has paved the way for a development of quantitative approaches in RNA computational biology, revealing unexpected connections between combinatorics and molecular biology. Using our strong background in enumerative combinatorics, we propose generic and efficient algorithms, both for sampling and counting structures using dynamic programming. These general techniques have been applied to study the sequence-structure relationship [77], the correction of pyrosequencing errors [71], and the efficient detection of multi-stable RNAs (riboswitches) [72], [73].

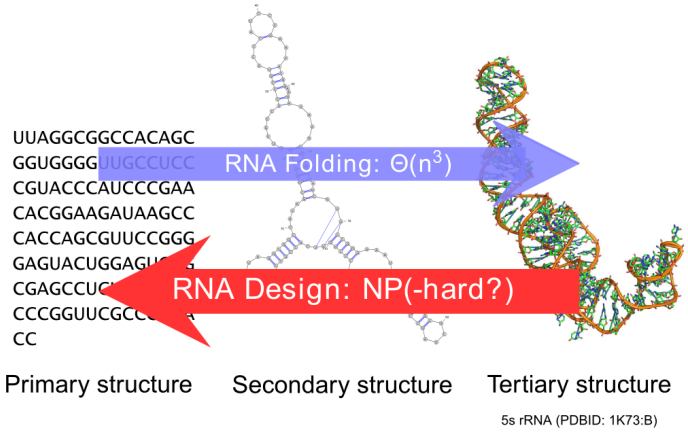

Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

3.1.2. RNA design.

Participants: Alain Denise, Vincent Le Gallic, Yann Ponty.

Joint project with S. Vialette (Marne-la-Vallée), J. Waldispühl (McGill) and Y. Zhang (Wuhan).

It is a natural pursue to build on our understanding of the secondary structure to construct artificial RNAs performing predetermined functions, ultimately targeting therapeutic and synthetic biology applications. Towards this goal, a key element is the design of RNA sequences that fold into a predetermined secondary structure, according to established energy models (inverse-folding problem). Quite surprisingly, and despite two decades of studies of the problem, the computational complexity of the inverse-folding problem is currently unknown.

Within our group, we offer a new methodology, based on weighted random generation [54] and multidimensional Boltzmann sampling, for this problem. Initially lifting the constraint of folding back into the target structure, we explored the random generation of sequences that are compatible with the target, using a probability distribution which favors exponentially sequences of high affinity towards the target. A simple posterior rejection step selects sequences that effectively fold back into the latter, resulting in a *global sampling* pipeline that showed comparable performances to its competitors based on local search [60].

3.1.3. Towards 3D modeling of large molecules

Participants: Alain Denise, Mélanie Boudard.

Joint project with D. Barth (Versailles) and J. Cohen (Paris-Sud).

The modeling of large RNA 3D structures, that is predicting the three-dimensional structure of a given RNA sequence, relies on two complementary approaches. The approach by homology is used when the structure of a sequence homologous to the sequence of interest has already been resolved experimentally. The main problem then is to calculate an alignment between the known structure and the sequence. The ab initio approach is required when no homologous structure is known for the sequence of interest (or for some parts of it). We work in both directions.

3.1.4. Statistical and robotics-inspired models for structure and dynamics

Participants: Julie Bernauer, Rasmus Fonseca.

Despite being able to correctly model small globular proteins, the computational structural biology community still craves for efficient force fields and scoring functions for prediction but also good sampling and dynamics strategies.

Our current and future efforts towards knowledge-based scoring function and ion location prediction have been described in 3.1.4.

Over the last two decades a strong connection between robotics and computational structural biology has emerged, in which internal coordinates of proteins are interpreted as a kinematic linkage with rotatable bonds as joints and corresponding groups of atoms as links [76], [51], [64], [63]. Initially, fragments in proteins limited to tens of residues were modeled as a kinematic linkage, but this approach has been extended to encompass (multi-domain) proteins [62]. For RNA, progress in this direction has been realized as well. A kinematics-based conformational sampling algorithm, KGS, for loops was recently developed [58], but it does not fully utilize the potential of a kinematic model. It breaks and recloses loops using six torsional degrees of freedom, which results in a finite number of solutions. The discrete nature of the solution set in the conformational space makes difficult an optimization of a target function with a gradient descent method. Our methods overcome this limitation by performing a conformational sampling and optimization in a co-dimension 6 subspace. Fragments remain closed, but these methods are limited to proteins. Our objective is to extend the approach proposed in [58], [76] to nucleic acids and protein/nucleic acid complexes with a view towards improving structure determination of nucleic acids and their complexes and in silico docking experiments of protein/RNA complexes. For that purpose, we have developed a generic strategy for differentiable statistical potentials [2], [74] that can be directly integrated in the procedure.

Results from in silico docking experiments will also directly benefit structure determination of complexes which, in turn, will provide structural insights in nucleic acid and protein/nucleic acid complexes. From the small proof-of-concept single chain protein implementation of the KGS strategy, we have developed a robust preliminary implementation that can handle RNA and will be further developed to account for multi-chain molecules. Rasmus Fonseca, post-doctoral scholar in the project is currently performing an extensive computational and biological validation.

3.2. Sequences

Participants: Alain Denise, Mireille Régnier, Yann Ponty, Jean-Marc Steyaert, Alice Héliou, Daria Iakovishina, Antoine Soulé.

String searching and pattern matching is a classical area in computer science, enhanced by potential applications to genomic sequences. In CPM/SPIRE community, a focus is given to general string algorithms and associated data structures with their theoretical complexity. Our group specialized in a formalization based on languages, weighted by a probabilistic model. Team members have a common expertise in enumeration and random generation of combinatorial sequences or structures, that are *admissible* according to some given constraints. A special attention is paid to the actual computability of formula or the efficiency of structures design, possibly to be reused in external software.

As a whole, motif detection in genomic sequences is a hot subject in computational biology that allows to address some key questions such as chromosome dynamics or annotation. This area is being renewed by high throughput data and assembly issues. New constraints, such as energy conditions, or sequencing errors and amplification bias that are technology dependent, must be introduced in the models. An other aim is to combine statistical sampling with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [66]. In general, in the future, our methods for sampling and sequence data analysis should be extended to take into account such constraints, that are continuously evolving.

3.2.1. Combinatorics of motifs

Participants: Mireille Régnier, Alice Héliou, Daria Iakovishina.

Besides applications [5] of analytic combinatorics to computational biology problems, the team addressed general combinatorial problems on words and fundamental issues on languages and data structures.

Molecular interactions often involve specific motifs. One may cite protein-DNA (cis-regulation), protein-protein (docking), RNA-RNA (miRNA, frameshift, circularisation). Motif detection combines an algorithmic search of potential sites and a significance assessment. Assessment significance requires a quantitative criterium. It is generally accepted that the p-value is a reliable tool that outperforms older criteria such as the z-score. AMIB develops a long term research on word combinatorics. In the recent years, a general scheme of derivation of analytic formula for the pvalue under different constraints (k-occurrence, first occurrence, overrepresentation in large sequences,...) has been provided. It relies on a representation of word overlaps in a graph [40]. Recursive equations to compute pvalues may be reduced to a traversal of that graph, leading to a linear algorithm. It allows for a derivation of pvalues, decreasing the space and time complexity of the generating function approach or previous probabilistic weighted automata.

In the mean time, continuous sequences of overlapping words, currently named *clumps* or *clusters* turn out to be crucial in random words counting. Notably, they play a fundamental role in the Chen-Stein method of compound Poisson approximation. A first characterization was proposed by Nicodème and al. and this work is currently extended.

This research area is widened by new problems arising from *de novo* genome assembly or re-assembly. For example, unique mappability of short reads strongly depends of the repetition of words. Although the average values for the length have been studied for long under different constraints, their distribution or profile remained unknown until the seminal paper [67] which provides formulae for binary tries. A collaboration has been started with LOB at Ecole Polytechnique to check these formulae on real data, namely Archae genomes (internship of D. Busatto-Gaston). This collaboration has been extended since LOB bought a sequencing machineand a co-advisd thesis (Alice Héliou) on circular RNA characterization hasjust started.

As a third example, one objective is to develop a model of errors, including a statistical model, that takes into account the quality of data for the different sequencing technologies, and their volume. This is the subject of an international collaboration with V. Makeev's lab (IoGene, Moscow) and MAGNOME project-team. Finally, Next Generation Sequencing open the way to the study of structural variants in the genome, as recently described in [48]. Defining a probabilistic model that takes into account main dependencies -such as the GC content- is a task of D. Iakovishina's thesis, to be defended in 2015, in a collaboration with V. Boeva (Curie Institute).

3.2.2. Random generation

Participants: Alain Denise, Yann Ponty.

Analytical methods may fail when both sequential and structural constraints of sequences are to be modelled or, more generally, when molecular *structures* such as RNA structures have to be handled. The random generation of combinatorial objects is a natural, alternative, framework to assess the significance of observed phenomena. General and efficient techniques have been developed over the last decades to draw objects uniformly at random from an abstract specification. However, in the context of biological sequences and structures, the uniformity assumption becomes unrealistic, and one has to consider non-uniform distributions in order to derive relevant estimates. Typically, context-free grammars can handle certain kinds of long-range interactions such as base pairings in secondary RNA structures.

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [55]: instead of formulating the problem as a classic optimization, this new approach uses statistical sampling within the space of solutions. Besides giving better, more robust, results, it allows for a fruitful adaptation of tools and algorithms derived in a purely combinatorial setting. Indeed, we have done significant and original progress in this area recently [68], [5], including combinatorial models for structures with pseudoknots. Our aim is to combine this paradigm with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [66].

Besides, our work on random generation is also applied in a different fields, namely software testing and model-checking, in a continuing collaboration with the Fortesse group at LRI [53], [65].

3.3. Geometry and machine learning for 3D interaction prediction

Participants: Julie Bernauer, Jean-Marc Steyaert, Christine Froidevaux, Adrien Guilhot-Gaudeffroy, Amélie Héliou.

The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. This is specially challenging as structure flexibility is key and multi-scale modelling [47], [57] and efficient code are essential [61].

Our project covers various aspects of biological macromolecule structure and interaction modelling and analysis. First protein structure prediction is addressed through combinatorics. The dynamics of these types of structures is also studied using statistical and robotics inspired strategies. Both provide a good starting point to perform 3D interaction modelling, accurate structure and dynamics being essential. Modelling is then raised to the cell level by studying large protein interaction networks and also the dynamics of molecular pathways.

Our group benefits from a good collaboration network, mainly at Stanford University (USA), HKUST (Hong-Kong) and McGill (Canada). The computational expertise in this field of computational structural biology is represented in a few large groups in the world (e.g. Pande lab at Stanford, Baker lab at U.Washington) that have both dry and wet labs. We also contributed to the CAPRI experiment organized by leading member of an international community we have been involved in for some time [56]. At Inria, our interest for structural biology is shared by the ABS project-team. A work by D. Ritchie in the ORPAILLEUR project-team (see [44] led to a joint publication with T. Bourquard and J. Azé. Our activities are however now more centered around protein-nucleic acid interactions, multi-scale analysis, robotics inspired strategies and machine learning than protein-protein interactions, algorithms and geometry. We also shared a common interest for large biomolecules and their dynamics with the NANO-D project team and their adaptative sampling strategy. As a whole, we contribute to the development of geometric and machine learning strategies for macromolecular docking.

3.3.1. Combinatorial models for the structure of proteins

Protein structure prediction has been and still is extensively studied. Computational approaches have shown interesting results for globular proteins but transmembrane proteins remain a difficult case.

Transmembrane beta-barrel proteins (TMB) account for 20 to 30% of identified proteins in a genome but, due to difficulties with standard experimental techniques, they are only 2% of the RCSB Protein Data Bank. As TMB perform many vital functions, the prediction of their structure is a challenge for life sciences, while the small number of known structures prohibits knowledge-based methods for structure prediction.

As barrel proteins are strongly structured objects, model based methodologies are an interesting alternative to these conventional methods. Jérome Waldisphül's thesis at LIX had opened this track for the common case where a protein folds respecting the order of the sequence, leaving a structure where each strand is bound to the preceding and succeeding ones. The matching constraints were expressed by a grammatical model, for which relatively simple dynamic programming schemes exist.

However, more sophisticated schemes are required when the arrangements of the strands along the barrel do not follow their order in the sequence, as it is the case for *Greek key* or *Jelly roll* motifs. The prediction algorithm may then be driven by a permutation on the order of the bonded strands. In his thesis [75], Van Du Tran developed a methodology for compiling a given permutation into a dynamic programming scheme that may predict the folding of sequences into the corresponding TMB secondary structure. Polynomial complexity upper bounds follow from the calculated DP scheme. Through tree decompositions of the graph that expresses constraints between strands in the barrel, better schemes were investigated in [75].

The efficiently obtained 3D structures provide a good model for further 3D and interaction analyses.

3.3.2. 3D interaction prediction

To better model complexes, various aspects of the scoring problem for protein-protein docking need being addressed [56]. It is also of great interest to introduce a hierarchical analysis of the original complex three-dimensional structures used for learning, obtained by clustering.

A protein-protein docking procedure traditionally consists in two successive tasks: a search algorithm generates a large number of candidate solutions, and then a scoring function is used to rank them in order to extract a native-like conformation. We demonstrated that, using Voronoi constructions and a defined set of parameters, we could optimize an accurate scoring function and interaction detection [46]. We also focused on developing other geometric constructions for that purpose: being related to the Voronoi construction, the Laguerre tessellation was expected to better represent the physico-chemical properties of the partners. It also allows a fast computation without losing the intrinsic properties of the biological objects. In [49], we compare both constructions. We also worked on introducing a hierarchical analysis of the original complex three-dimensional structures used for learning, obtained by clustering. Using this clustering model, in combination with a strong emphasis on the design of efficient complex filters collaborative filtering, we can optimize the scoring functions and get more accurate solutions [50].

These techniques have been extended to the analysis of protein-nucleic acid complexes : developments and tests are performed by A. Guilhot (See figure 2) in his PhD thesis.

3.4. Data Integration

Participants: Christine Froidevaux, Alain Denise, Sarah Cohen-Boulakia, Bryan Brancotte, Jiuqiang Chen.

Faced with the inherent features of biological and biomedical data, researchers from the database and artificial intelligence communities have joined together to form a community dedicated to the study of the specific problems posed by integrating life sciences data. With the deluge of new sequenced genome sequences and the amount of data produced by high-throughput approaches, the need to cross and compare massive and heterogeneous data is more important than ever to improve functional annotation and design biological networks. Challenges are numerous. One may cite the need to provide support to scientists to perform and share complex and reproducible complex biological analyses. A special attention is paid to the more specific domain of scientific workflows management and ranking biological data. One aims at exploring the relationships between those two domains, from the investigation of various specific problems posed by ranking scientific workflows to the problem of considering consensus workflows.

3.4.1. Designing and Comparing Scientific workflows

Participants: Sarah Cohen-Boulakia, Christine Froidevaux, Jiuqiang Chen.

Scientific workflows management systems are increasingly used to specify and manage bioinformatics experiments. Their programming model appeals to bioinformaticians, who use them to easily specify complex data processing pipelines. Such a model is underpinned by a graph structure, where nodes represent bioinformatics tasks and links represent the dataflow. As underlined both in a study and a review of existing approaches, the complexity of such graph structures is increasing over time, making them more difficult to share and reuse.

One of the major current challenges is thus to provide means to reduce the structural complexity of workflows while ensuring that any structural transformation will not have any impact on the executions of the transformed workflows, that is, preserving *provenance*.

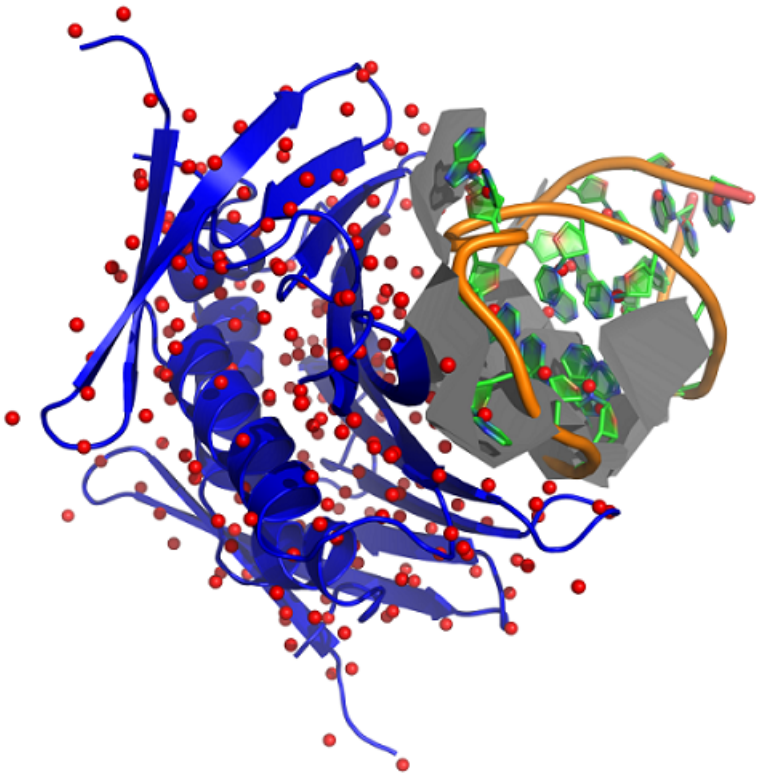

Figure 2. Coarse-grained representation and Voronoi interface model of a PP7 coat protein bound to an RNA hairpin (PDB code 2qux). The Voronoi model captures the features of the interactions such as stacking, even at the coarse-grained level.

3.4.2. Ranking biological data

Participants: Alain Denise, Sarah Cohen-Boulakia, Bryan Brancotte.

We are addressing the increase of the number of resources available. The BIOGUIDE project aim at helping user navigation in the maze of available biological sources. More recently, a second problem was tackled: the number of answers returned by even one single queried biological resource may be too large for the user to deal with. We have provided solutions for ranking biological data. The main difficulty lies in considering various ranking criteria (recent data first, popular data first, curated data first...). Many approaches combine ranking criteria to design a ranking function, possibly leading to arbitrary choices made in the way of combining the ranking criteria. Instead, in collaboration with the University of Montreal, we have proposed to follow a *median ranking approach* named BIOCONSERT (for generating Biological Consensus ranking with ties): considering as many rankings as they are ranking criteria for the same data set, and providing a consensus ranking that minimizes the disagreements between the input rankings. We have shown the benefit of using median ranking in several biological settings.

Additionally, in a close collaboration with the Institut Curie, we have also developed the GENEVALORIZATION tool that ranks a list of genes of interest given as input with respect to a set of keywords representing the context of study. Here the single ranking criterion considered for each gene is the number of publications in PubMed co-citing the gene name and the keywords. The tool is able to make use of the MeSH taxonomy when considering the keywords and the dictionary of gene names and aliases for the gene names.

3.5. Systems Biology

Participants: Patrick Amar, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Loic Paulevé, Sabine Pérès, Jean-Marc Steyaert, Erwan Bigan, Adrien Rougny.

Systems Biology involves the systematic study of complex interactions in biological systems using an integrative approach. The goal is to find new emergent properties that may arise from the systemic view in order to understand the wide variety of processes that happen in a biological system. Systems Biology activity can be seen as a cycle composed of theory, computational modelling to propose a hypothesis about a biological process, experimental validation, and use of the experimental results to refine or invalidate the computational model (or even the whole theory). During the past five years, new questions and research domains have been identified, and some members of the team have reoriented a part of their activities on these questions.

Three main types of problems have been studied: metabolic networks, signaling networks and more recently synthetic biology. Networks - have become popular since many crucial problems, coming from biology, medecine, pharmacology, are nowadays stated in these terms: a great number of them are issued from the cancer phenomenom and the will to enhance our understanding in order to propose more efficient therapeutic issues. Metabolism has received the major attention since it concerns a large variety of topics and several methods that have been proposed. Depending on the nature of the biological problem, several methods can be used : discrete deterministic, stochastic, combinatorial, up to continuous differential. Also, the recent rise of synthetic biology proposes similar challenges aiming at improving the production of energy by means of biological systems or at getting more efficient medicamental treatments, for instance.

3.5.1. Topological analysis of metabolic networks

Participant: Sabine Pérès.

Elementary flux mode analysis is a powerful tool for the theoretical study of simple metabolic networks. However, when the networks are complex, the determination of elementary flux modes leads to a combinatorial explosion of their number which prevents from drawing simple conclusions from their analysis. Since the concept of elementary flux mode analysis was introduced in 1994, there has been an important and ongoing effort to develop more efficient algorithms. However, these methods share a common bottleneck: they enumerate all the elementary flux modes which make the computation impossible when the metabolic network is large and only few works try to search only elementary flux mode with specific properties. We have shown that enumerating all the elementary flux modes is not necessary in many cases and it is possible

to directly query the network instead with an appropriate tool. For ensuring a good query time, we have relied on a state of the art SAT solver, working on a propositional encoding of elementary flux mode, and enriched with a simple SMT-like solver ensuring elementary flux mode consistency with stoichiometric constraints. We have illustrated our new framework by providing experimental evidences of almost immediate answer times on a non trivial metabolic network [45], [70].

3.5.2. Signaling networks

Participants: Sarah Cohen-Boulakia, Christine Froidevaux, Adrien Rougny.

Signaling pathways involving G protein-coupled receptors (GPCR) are excellent targets in pharmacogenomics research. Large amounts of experiments are available in this context while globally interpreting all the experimental data remains a very challenging task for biologists. Our goal is to help the understanding of signaling pathways involving (GPCR) and to provide means to semi-automatically construct the signaling networks.

We have introduced a logic-based method to infer molecular networks and show how it allows inferring signaling networks from the design of a knowledge base. Provenance of inferred data has been carefully collected, allowing quality evaluation. Our method (i) takes into account various kinds of biological experiments and their origin; (ii) mimics the scientist's reasoning within a first-order logic setting; (iii) specifies precisely the kind of interaction between the molecules; (iv) provides the user with the provenance of each interaction; (v) automatically builds and draws the inferred network [43].

Observe that a logic-based formalisation is used as in some works carried out in INRIA team DYLISS. AMIB aim is different, as the design of the network lies on a knowledge-based system describing experimental facts and ontological relationships on backgound knowledge, together with a set of generic and expressive rules, that mimick the expert's reasoning.

This is a collaboration with A. Poupon (INRA-BIOS, Tours) that was supported by an INRA-INRIA starting grant in 2011-2012.

3.5.3. Modelling and Simulation

Participants: Patrick Amar, Sarah Cohen-Boulakia, Loic Paulevé, Jean-Marc Steyaert, Erwan Bigan.

A great number of methods have been proposed for the study of the behavior of large biological systems. The first one is based on a discrete and direct simulation of the various interactions between the reactants using an entity-centered approach; the second one implements a very efficient variant of the Gillespie stochastic algorithm that can be mixed with the entity-centered method to get the best of both worlds; the third one uses differential equations automatically generated from the set of reactions defining the network.

These three methods have been implemented in an integrated tool, the HSIM system [41]. It mimics the interactions of biomolecules in an environment modelling the membranes and compartments found in real cells. It has been applied to the modelling of the circadian clock of the cyanobacterium, and we have shown pertinent results regarding the spontaneous appearance of oscillations and the factors governing their period [42].

3.5.3.1. Synthetic biology

Synthetic biology begins to be a very popular domain of research. Genetic engineering is a good example of synthetic biology, organisms are artificially modified to boost the production of compounds that might be used in the medical or industrial domains. We have been focused on using synthetic biology for medical diagnostic purposes. In a collaboration with the SYSDIAGLab (UMR 3145) at Montpellier, P. Amar participates at the COMPUBIOTIC project. The goal is to design, test and build an artificial embedded biological nano-computer in order to detect the biological markers of some human pathologies (colorectal cancer, diabetic nephropathy, etc.). This nano-computer is a small vesicle containing specific enzymes and membrane receptors. These components are chosen in a way that their interactions can sense and report the presence in the environment of molecules involved in the human pathologies tageted. We plan to design a dedicated software suite to help the design and validation of this artificial nano-computer. HSIM is used to help the design and to test qualitatively and quantitatively this *"biological computer"* before *in vitro*.

3.5.3.2. Evaluating metabolic networks

It is now well established in the medical world that the metabolism of organs depends crucially of the way the calls consume oxygene, glucose and the various metabolites that allow them to grow and duplicate. A particular variety of cells, tumour cells, is of major interest. In collaboration with L. Schwartz (AP-HP) and biologists from INSERM-INRA Clermont-Theix we have started a project aiming at identifying the important points in the metabolic machinery that command the changes in behaviour. The main difficulties come from the fact that biologists have listed dozens of concurrent cycles that can be activated alternatively or simultaneously, and that the dynamic characteristics of the chemical reactions are not known accurately.

Given the set of biochemical reactions that describe a metabolic function (e.g. glycolysis, phospholipids' synthesis, etc.) we translate them into a set of o.d.e's whose general form is most often of the Michaelis-Menten type but whose coefficients are usuall very badly determined. The challenge is therefore to extract information as to the system's behavior while making reasonable asumptions on the ranges of values of the parameters. It is sometimes possible to prove mathematically the global stability, but it is also possible to establish it locally in large subdomains by means of simulations.Our program `Mpas` (Metabolic Pathway Analyser Software) renders the translation in terms of a systems of o.d.e's automatic, leading to easy, almost automatic simulations. Furthermore we have developed a method of systematic analysis of the systems in order to characterize those reactants which determine the possible behaviors: usually they are enzymes whose high or low concentrations force the activation of one of the possible branches of the metabolic pathways. A first set of situations has been validated with a research INSERM-INRA team based in Clermont-Ferrand. In her PhD thesis, defended in 2011, M. Behzadi proved mathematically the decisive influence of the enzyme PEMT on the Choline/Ethylamine cycles.

3.5.3.3. Comparison of Metabolic Networks

We study the interest of *fungi* for biomass transformation. Cellulose, hemicellulose and lignin are the main components of plant biomass. Their transformation represent a key energy challenges of the 21st century and should eventually allow the production of high value new compounds, such as wood or liquid biofuels (gas or bioethanol). Among the boring organisms, two groups of fungi differ in how they destroy the wood compounds. Analysing new fungi genomes can allow the discover of new species of high interest for bio-transformation. For a better understanding of how the fungal enzymes facilitates degradation of plant biomass, we conduct a large-scale analysis of the metabolism of fungi. Machine learning approaches such like hierarchical rules prediction are being studied to find new enzymes allowing the transformation of biomass. The KEGG database http://www.genome.jp/kegg/ contains pathways related to fungi and other species. By analysing these known pathways with rules mining approaches, we aim to predict new enzymes activities.

4. New Software and Platforms

4.1. Cartaj

Participant: Alain Denise [correspondant].

CARTAJ is a software that automatically predicts the topological family of three-way junctions in RNA molecules, from their secondary structure only : :the sequence and the canonical Watson–Crick pairings. The Cartaj software http://cartaj.lri.fr that implements our method can be used online. It is also meant for being part of RNA modelling softwares and platforms. The methodology and the results of CARTAJ are presented in [59]. More than 300 visits since its release in January 2012.

4.2. DiMoVo

Participant: Julie Bernauer [correspondant].

DIMOVO, *DIscriminate between Multimers and MOnomers by VOronoi tessellation* : Knowing the oligomeric state of a protein is necessary to understand its function. his tool, accessible as a webserver and still used and maintained, provides a reliable discrimination function to obtain the most favorable state of proteins.

Availability : released in 2008.

4.3. VorScore

Participant: Julie Bernauer [correspondant].

VORSCORE, *Voronoi Scoring Function Server* : Scoring is a crucial part of a protein-protein procedure and having a quantitave function to evaluate conformations is mandatory. This server provides access to a geometric knowledge-based evaluation function. It is still maintained and widely used. See Bernauer et al., Bioinformatics, 2007 23(5):555-562 for further details.

4.4. ConQuR-Bio

Participants: Bryan Brancotte, Sarah Cohen-Boulakia [correspondant], Alain Denise.

ConQuR-Bio assists scientists when they query public biological databases. Various reformulations of the user query are generated using medical terminologies (MeSH, OMIM, ...). Such alternative reformulations are then used to rank the query results using a new consensus ranking strategy. The originality of our approach thus lies in using consensus ranking techniques within the context of query reformulation. The ConQuR-Bio system is able to query the Entrez-Gene NCBI database. The benefit of using ConQuR-Bio compared to what is currently provided to users has been demonstrated on a set of biomedical queries.
Availability : http://conqur-bio.lri.fr/

4.5. VARNA (Visualization Application for RNA)

Participants: Yann Ponty [correspondant], Alain Denise.

A lightweight Java Applet dedicated to the quick drawing of an RNA secondary structure. VARNA is open-source and distributed under the terms of the GNU GPL license. Automatically scales up and down to make the most out of a limited space. Can draw multiple structures simultaneously. Accepts a wide range of documented and illustrated options, and offers editing interactions. Exports the final diagrams in various file formats (svg,eps,jpeg,png,xfig) [52]...
VARNA currently ships in its 3.9 version, and consists in ∼50 000 lines of code in ∼250 classes.
Availability : Distributed at http://varna.lri.fr since 2009 under the GPL v3 license.
Impact: Downloaded ∼15k times and cited by ∼250 research manuscripts (source: Google Scholar).

4.6. GenRGenS (GENeration of Random GENomic Sequences)

Participants: Yann Ponty [correspondant], Alain Denise.

A software dedicated to the random generation of sequences. Supports different lasses of models, including weighted context-free grammars, Markov models, PROSITE patterns... [69] GENRGENS currently ships in its 2.0 version, and consists in ∼25 000 lines of code in ∼120 Java classes.
Availability : Distributed at http://www.lri.fr/~genrgens/ since 2006 under the terms of the GPL v3 license.
Impact: Downloaded ∼5k times and cited by ∼60 times (source: Google Scholar).

4.7. GeneValorization

Participants: Bryan Brancotte, Sarah Cohen-Boulakia [correspondant].

High-throughput technologies provide fundamental informations concerning thousands of genes. Most of the current biological research laboratories daily use one or more of these technologies and identify lists of genes. Understanding the results obtained includes accessing to the latest publications concerning individual or multiple genes. Faced to the exponential growth of publications avaliable, this task is becoming particularly difficult to achieve.

Here, we introduce a web-based Java application tool named GeneValorization which aims at making the most of the text-mining effort done downstream to all high throughput technology assays. Regular users come from the Curie Institute, but also the EBI.
Impact : 925 distinct international users have used GeneValorization and about a hundred use it on a regular basis. The tool is on average used once to twice every day.
Availability : it is available at http://bioguide-project.net/gv with Inter Deposit Digital Number (*depot APP*, June 2013).

4.8. HSIM

Participant: Patrick Amar [correspondant].

HSIM (Hyperstructure Simulator) is a simulation tool for studying the dynamics of biochemical processes in a virtual bacteria. The model is given using a language based on probabilistic rewriting rules that mimics the reactions between biochemical species. HSIM is a stochastic automaton that implements an entity-centered model of objects. This kind of modelling approach is an attractive alternative to differential equations for studying the diffusion and interaction of the many different enzymes and metabolites in cells which may be present in either small or large numbers.

The new version of HSIM includes a Stochastic Simulation Algorithm *a la* Gillespie that can be used with the same model in a standalone way or in a mixed way with the entity-centered algorithm. This new version offers also the possibility to export the model in SciLab for a ODE integration. Last, HSIM can export the differential equations system, equivalent to the model, to LaTeX for pretty-printing.

This software is freely available at http://www.lri.fr/~pa/Hsim; A compiled version is available for the Windows, Linux and MacOSX operating systems.

4.9. Pint

Participant: Loïc Paulevé [correspondant].

PINT provides several command-line tools to model, simulate, and analyse the dynamics of automata networks. Its main application domain is systems biology for modelling and analysis of very large interaction networks. Besides a textual language for specifying networks and standard stochastic simulation algorithms, PINT implements static analysis for analysing and controlling the transient reachability. In particular, PINT provides the computation of cut sets for transient reachability, that gives sets of key automata states, whose mutation would prevent the concerned reachability to occur.

PINT has been applied to extremely large biological networks, from 100 to 10,000 interacting components, demonstrating its scalability and potential to handle full databases of interactions.

PINT is distributed under the CeCiLL licence, and is available at http://loicpauleve.name/pint.

5. New Results

5.1. RNA

To mitigate the current absence of a selective scientific event dedicated to RNA computational biology, impeding the dissemination of recent methodological results, AMIB members have participated in the creation of the *Computational Methods for Structural RNAs* workshops (CMSR'14). This first installment of the event was hosted in Strasbourg as a workshop of the 2014 edition of European Conference on Computational Biology. Its proceedings were published by McGill University [33], and extended versions of selected articles were invited to appear in the *Journal of Computational Biology*.

5.1.1. RNA visualization

The field of RNA visualization is now rich with multiple tools that accommodate different needs, arising from a variety of application contexts. In order to help end-users navigate through the jungle of available options, Y. Ponty and F. Leclerc (IGM, Univ. Paris-Sud) have contributed a review of existing tools, and illustrate their usage to address a collection of typical use-cases [35].

5.1.2. RNA design and structures

The past couple of years have seen the multiplication of heuristic or exponential time algorithms for the RNA design problem. This situation motivates a survey, which s currently lacking, that would focus on the relative merits of existing algorithms, and assess their applicability towards the typical goals of synthetic biology. Such an objective evaluation is at the core of the PhD project of Vincent Le Gallic, which was started in September 2014.

With Antoine Soulé, a PhD student of J-M Steyaert and J. Waldispühl (McGill), a comparative study of the various softwares for the inverse RNA folding problem is under revision and a new version of RNAMUTANT in the langage GAP-L with enrichment has been designed.

Besides, we have published a general survey on RNA structure comparison [9].

5.1.3. RNA splicing regulation

RNA splicing is a modification of the nascent pre-messenger RNA (pre-mRNA) transcript in which introns are removed and exons are joined. The U2AF heterodimer protein has been well studied for its role in defining functional 3' splice sites in pre-mRNA splicing, but multiple critical problems are still outstanding, including the functional impact of their cancer-associated mutations. In collaboration with Xiang-Dong Fu's groups in San Diego and Wuhan, , through genome-wide analysis of U2AF-RNA interactions, we reported in [16] that U2AF has the capacity to define 88% of functional 3' splice sites in the human genome. Numerous U2AF binding events also occur in other genomic locations, and metagene and minigene analysis suggests that upstream intronic binding events interfere with the immediate downstream 3' splice site associated with either the alternative exon to cause exon skipping or competing constitutive exon to induce inclusion of the alternative exon.

5.1.4. RNA 3D structure modelling

Conformational diversity for RNA ensemble analyses is often provided by sophisticated molecular dynamics simulations. Long trajectories with specialized force fields on dedicated supercomputers are required to adequately sample conformational space, limiting ensemble analyses to modestly-sized RNA molecules. To avoid these limitations, we developed an efficient conformational sampling procedure, Kino-geometric sampling for RNA (KGSrna), which can report on ensembles of RNA molecular conformations orders of magnitude faster than MD simulations. In the KGSrna model, the RNA molecule is represented with rotatable, single bonds as degrees-of-freedom and groups of atoms as rigid bodies. In this representation, non-covalent bonds form distance constraints, which create nested, closed cycles in a rooted spanning tree. Torsional degrees-of-freedom in a closed ring demand carefully coordinated changes to avoid breaking the non-covalent bond, which greatly reduces the conformational flexibility. The reduced flexibility from a network of nested, closed rings consequently deforms the biomolecule along preferred directions on the conformational landscape. This new procedures projects degrees-of-freedom onto a lower-dimensional subspace of the conformation space, in which the geometries of the non-covalent bonds are maintained exactly under conformational perturbation. The dimensionality reduction additionally enables efficient exploration of conformational space and reduces the risk of overfitting sparse experimental data. Kinogeometric sampling of 3D RNA models can recover the conformational landscape encoded by proton chemical shifts in solution and is thus of great help to interpret NMR experimental data [11]. The computational efficiency of this approach, combined to its inherent parallel nature could also be adapted to model large assemblies on parallel platforms.

Our expertise was also essential in modelling junction of the RNA structure of a large biomolecule of interest, the tmRNA so as to study its interaction with the SmpB protein. Results obtained in collaboration with experiementalists, mainly P. Vachette at IBBMC and S. Nonin-Lecomte at the LCRB were made available in [15].

5.2. Sequences

5.2.1. Random generation

In collaboration with the Simon Fraser University (Vancouver, Canada), we have explored a random generation strategy, under a Boltzmann distribution, to assess the robustness of predicted adjacencies in ancestral genomes using a parsimony-based approach. The sampling algorithm was used to estimate the Boltzmann probability of ancestral adjacencies, which was then used as a filter to weed out unsupported predictions, leading to the resolution of a large number of syntenic inconsistencies [23].

5.2.2. Combinatorics of motifs

An algorithm for pvalue computation has been proposed in [40] that takes into account a Hiddden Markov Model and an implementation, SUFPREF, has been realized (http://server2.lpm.org.ru/bio).

Combinatorics of clumps have been extensively studied, leading to the definition of the so-called *canonic clumps*. It is shown in [26] that they contain the necessary information needed to calculate, approximate, and study probabilities of occurrences and asymptotics. This motivates the development of a *clump automaton*. It allows for a derivation of pvalues, decreasing the space and time complexity of the generating function approach or previous weighted automata. An extension to degenerate patterns is currently realized and implemented in a collaboration with J. Holub (Praha U.) and E. Furletova (IMPB).

During her master thesis at King's College, A. Héliou and collaborators designed the first linear-time and linear-space algorithm for computing all minimal absent words based on the suffix array [6]. In a typical application, one would be interested in computing minimal absent words to compare and study genomes in linear time by considering this negative information.

In a collaboration with AlFarabi University, where M. Régnier acts as a foreign co-advisor), word statistics were used to identify mRNA targets for miRNAs involved in various cancers [7].

5.2.3. Prediction and functional annotation of ortholog groups of proteins

In comparative genomics, orthologs are used to transfer annotation from genes already characterized to newly sequenced genomes. Many methods have been developed for finding orthologs in sets of genomes. However, the application of different methods on the same proteome set can lead to distinct orthology predictions.

In [38], [14] we developed a method based on a meta-approach that is able to combine the results of several methods for orthologous group prediction. The purpose of this method is to produce better quality results by using the overlapping results obtained from several individual orthologous gene prediction procedures. Our method proceeds in two steps. The first aims to construct seeds for groups of orthologous genes; these seeds correspond to the exact overlaps between the results of all or several methods. In the second step, these seed groups are expanded by using HMM profiles.

We evaluated our method on two standard reference benchmarks, OrthoBench and Orthology Benchmark Service. Our method presents a higher level of accurately predicted groups than the individual input methods of orthologous group prediction. Moreover, our method increases the number of annotated orthologous pairs without decreasing the annotation quality compared to twelve state-of-the-art methods.

5.3. 3D Modelling and Interactions

5.3.1. Transmembrane proteins

Transmembrane beta-barrel proteins (TMB) account for 20 to 30% of identified proteins in a genome but, due to difficulties with standard experimental techniques, they are only 2% of the RCSB Protein Data Bank. As

TMB perform many vital functions, the prediction of their structure is a challenge for life sciences, while the small number of known structures prohibits knowledge-based methods for structure prediction. We study and design algorithmic solutions addressing the secondary structure, an abstraction of the 3D conformation of a molecule, that only retains the contacts between its residues. As TMBs are strongly structured objects, model based methodologies [18] are an interesting alternative to conventional methods. The efficiently obtained 3D structures provide a good model for further 3D and interaction analyses.

5.3.2. 3D Interaction prediction

While protein-RNA complexes provide a wide range of essential functions in the cell, their atomic experimental structure solving is even more difficult than for proteins. Protein-RNA complexes provide a wide range of essential functions in the cell. Docking approaches that have been developed for proteins are often challenging to adapt for RNA because of its inherent flexibility and the structural data available being relatively scarce.We adapted the reference RosettaDock protocol for protein-RNA complexes both at the nucleotide and atomic levels. Using a genetic algorithm-based strategy, and a non-redundant protein-RNA dataset, we derived a RosettaDock scoring scheme able not only to discriminate but also score efficiently docking decoys. The approach proved to be both efficient and robust for generating and identifying suitable structures when applied to two protein-RNA docking benchmarks in both bound and unbound settings. It also compares well to existing strategies. This is the first approach that currently offers a multi-level optimized scoring approach integrated in a full docking suite, leading the way to adaptive fully flexible strategies [28], [12]. This work is part of the PhD thesis of Adrien Guilhot-Gaudeffroy. While the previously described approaches perform well in a rigid or semi-flexible docking setting, the generation of putative conformations for flexible molecules (sampling) is still a difficult question that has to be addressed in a multi-scale setting involving new algorithms. Docking these sampled conformations will also certainly require improvement in clustering approaches.

5.4. Data Integration

With the increasing popularity of scientific workflows, public and private repositories are gaining importance as a means to share, find, and reuse such workflows.

As the sizes of these repositories grow, methods to compare the scientific workflows stored in them become a necessity, for instance, to allow duplicate detection or similarity search. Scientific workflows are complex objects, and their comparison entails a number of distinct steps from comparing atomic elements to comparison of the workflows as a whole. Various studies have implemented methods for scientific workflow comparison and came up with often contradicting conclusions upon which algorithms work best. Comparing these results is cumbersome, as the original studies mixed different approaches for different steps and used different evaluation data and metrics.

In collaboration with members of the University of Humboldt (Berlin), we first contribute to the field [17] by (i) comparing in isolation different approaches taken at each step of scientific workflow comparison, reporting on an number of unexpected findings, (ii) investigating how these can best be combined into aggregated measures, and (iii) making available a gold standard of over 2000 similarity ratings contributed by 15 workflow experts on a corpus of 1500 workflows and re-implementations of all methods we evaluated. In this context, we have designed new approaches based on consensus ranking [21] to provide a consensus of the experts' answers.

Then, with members of the University of Pennsylvania, we have presented a novel and intuitive workflow similarity measure that is based on layer decomposition [27] (designed during the month SCB spent at UPenn). Layer decomposition accounts for the directed dataflow underlying scientific workflows, a property which has not been adequately considered in previous methods. We comparatively evaluate our algorithm using our gold standard and show that it a) delivers the best results for similarity search, b) has a much lower runtime than other, often highly complex competitors in structure-aware workflow comparison, and c) can be stacked easily with even faster, structure-agnostic approaches to further reduce runtime while retaining result quality.

Another way to make scientific workflows easier to reuse is to reduce their structural complexity to make them easier to apprehend. In particular, we have continued to work in collaboration with the University of Manchester on DistillFlow, an approach to remove the structural redundancy in workflows. Our contribution is four fold. Firstly, we identify a set of anti-patterns that contribute to the structural workflow complexity. Secondly, we design a series of refactoring transformations to replace each anti-pattern by a new semantically-equivalent pattern with less redundancy and simplified structure. Thirdly, we introduce a distilling algorithm that takes in a workflow and produces a distilled semantically-equivalent workflow [8]. Lastly, we provide an implementation of our refactoring approach (dedicated demo published [24]) that we evaluate on both the public Taverna workflows and on a private collection of workflows from the BioVel project. On going work includes extending the list of anti-patterns to be considered and identifying *good patterns*, that is, patterns which are easy to maintain and have systematically been able to be executed. This has been done in the context of the master internship of Stéphanie Kamgnia Wonkap [37]. First results obtained are promising.

5.5. Systems Biology

5.5.1. Analyzing SBGN-AF Networks Using Normal Logic Programs

A wide variety of signaling networks are available in the literature or in databases under the form of influence graphs. In order to understand the systems underlying these networks and to modify them for a medical purpose, it is necessary to understand their dynamics. Consequently, a variety of modelling techniques for these networks have been developped. In particular, it is possible to model their dynamical behavior with Boolean networks. The construction of these Boolean networks starting from influence graphs requires a parametrization of some Boolean functions. This task is most often realized by interpreting experimental results, that can be hard to obtain.

We introduced a method that allows to model any influence graph expressed in the Systems Biology Graphical Notation Activity Flow language (SBGN-AF) under the form of a Boolean network [32], [29]. The parametrization does not rely on any experimental results but on general principles that govern the dynamics of signaling networks. Together with the translation of a SBGN-AF influence graph into predicates, these general principles expressed under the form of logic rules form a first-order normal logic program (NLP) equivalent to a Boolean network. We show that the trajectories as well as the steady-state of any SBGN-AF network can be obtained by computing the orbits and the supported models of its corresponding NLP, respectively.

5.5.2. Scalable methods for analysing dynamics of automata networks

In collaboration with T. Chatain, S. Haar, S. Schwoon, and L. Jezeguel (INRIA MEXICO), we explored new techniques for computing the reachable attractors in automata networks using Petri net unfoldings [22]. Attractors of network dynamics represent the long-term behaviours of the modelled system. Their characterization is therefore crucial for understanding the response and differentiation capabilities of a dynamical system. In the scope of qualitative models of interaction networks, the computation of attractors reachable from a given state of the network faces combinatorial issues due to the state space explosion. Our new algorithm relies on Petri net unfoldings that can be used to compute a compact representation of the dynamics, in particular by exploiting the concurrency of the transitions in order to remove redundant sequences of transitions. We illustrate the applicability of the algorithm with Petri net models of cell signalling and regulation networks, Boolean and multi-valued. The proposed approach aims at being complementary to existing methods for deriving the attractors of Boolean models, while being generic since it actually applies to any safe Petri net.

In collaboration with M. Folschette, M. Magnin, O. Roux (IRCCYN, Nantes), and K. Inoue (NII, Tokyo), we developed a framework for identifying classical Boolean or discrete networks models from Proces Hitting (PH) models [10]. The PH allows to model non-deterministic cooperations between interacting components, and we have shown that the dynamics of a single PH can embed (include) the dynamics of multiple discrete networks, where transitions functions are deterministic. Hence, if a behaviour is shown impossible at the PH model, it is necessary impossible in any included discrete models. Such kind of analysis is relevant in systems biology, where the cooperations between components are often under-determined and the enumeration of all

compatible discrete models is intractable: our framework allows to reason on the dynamics of a single abstract model.

Finally, a chapter summarizing the recent advances on static analysis for dynamics of large biological networks has been published as part of the *Logical Modeling of Biological Systems* handbook [30].

6. Partnerships and Cooperations

6.1. National Initiatives

6.1.1. ANR

A. Denise is involved in the NSD-NGD ANR project 2010-2014. Y. Ponty was involved in the MAGNUM ANR project (BLAN program, 12/2010–12/2014).

6.1.2. PEPS

Ch. Froidevaux was responsible at LRI for the CNRS-INSERM-INRIA PEPS grant *Identification of metabolic capabilities of fungi by comparative genomic* involving IGM, Paris-Sud and UMR GV, CNRS.

6.1.3. FRM

Fondation pour la Recherche Medicale – *Analyse Bio-informatique pour la recherche en Biologie* program

- Approche comparatives haut-débit pour la modelisation de l'architecture 3D des ARN à partir de données experimentales
- 2015–2018
- Y. Ponty, A. Denise
- B. Sargueil (Paris V – Experimental partner), J. Waldispuhl

6.2. European Initiatives

6.2.1. Collaborations in European Programs, except FP7 & H2020

ANR International program

- Fast and efficient sampling of structures in RNA folding landscapes
- RNALands (ANR-14-CE34-0011)
- 01/10/2014-30/09/2018
- Y. Ponty, A. Denise, M. Regnier
- EPI BONSAI/INRIA Inria Lille - Nord Europe, Vienna University (Austria)

6.3. International Initiatives

- Capes Biologie systémique du cancer (051/2013) porté par Sandro José de Souza (Univ. Federal do Rio Grande do Norte, Brésil)
- Sabine Peres
- 2014-2018

6.3.1. Inria Associate Teams

6.3.1.1. ITSNAP

Title: Intelligent Techniques for Structure of Nucleic Acids and Proteins

International Partner (Institution - Laboratory - Researcher):

Stanford University (ÉTATS-UNIS)

Duration: 2009 - 2014

See also: http://pages.saclay.inria.fr/julie.bernauer/EA_ITSNAP/

The ITSNAP Associated Team project is dedicated to the computational study of RNA 3D structure and interactions. By developing new molecular hierarchical models for knowledge-based and machine learning techniques, we can provide new insights on the biologically important structural features of RNA and its dynamics. This knowledge of RNA molecules is key in understanding and predicting the function of current and future therapeutic targets.

6.3.2. Inria International Partners

6.3.2.1. Declared Inria International Partners

Title: CARNAGE: Combinatorics of Assembly and RNA in GEnomes

International Partner (Institution - Laboratory - Researcher):

State Research Institute of Genetics and Selection of Industrial Microorganisms (Russia (Russian Federation)) - Bioinformatics laboratory - V. Makeev and Mireille Régnier

Duration: 2012- 2014

See also: https://team.inria.fr/amib/carnage

CARNAGE addresses two main issues on genomic sequences, by combinatorial methods.

Fast development of high throughput technologies has generated a new challenge for computational biology. The recently appeared competing technologies each promise dramatic breakthroughs in both biology and medicine. At the same time the main bottlenecks in applications are the computational analysis of experimental data. The sheer amount of this data as well as the throughput of the experimental dataflow represent a serious challenge to hardware and especially software. We aim at bridging some gaps between the new "next generation"sequencing technologies, and the current state of the art in computational techniques for whole genome comparison. Our focus is on combinatorial analysis for NGS data assembly, interspecies chromosomal comparison, and definition of standard pipelines for routine large scale comparison.

This project also addresses combinatorics of RNA and the prediction of RNA structures, with their possible interactions.

6.3.2.2. Informal International Partners

Polytechnique/UPSud and McGill/U. Montréal

Program: CFQCU

Title: Réseau franco-québecois de recherche sur l'ARN

Inria principal investigator: Jean-Marc Steyaert

International Partner (Institution - Laboratory - Researcher):

Mc Gill and Université de Montréal (Canada)

Computer Science Department

Jérôme Waldispühl

Duration: 2012 - 2014

Résumé : The partners have developped complementary expertise on RNA : bioinformatics, combinatorics and algorithms. machine learning, physics and genomics. Methodologies will be developed that combine theoretical simulations and new (high throughput) experimental data. A common high level training at Master and PhD level is organized.

6.3.3. Participation In other International Programs

Henry van den Bedem and J. Bernauer presented their work at the Inria BIS 2014 Workshop in Paris https://project.inria.fr/inria-siliconvalley/workshops/bis2014/.

6.4. International Research Visitors

6.4.1. Visits of International Scientists

J. Holub

Subject: Word automata

Institution: Praha University (Czech Republic)

E. Furletova

Subject: word enumeration

Institution: Institute of Mathematical Problems in Biology (Russia)

6.4.1.1. Internships

Jan Lin Chan

Subject: Exceptional words in *Archae* genomes

Date: 01/06/2014 - 11/08/2014

Institution: NUS (Singapour)

Funding: INRIA

Supervision: M. Régnier

Damien Busatto-Gaston

Subject: de Bruijn graphs and assembly

Date: 01/06/2014 - 14/07/2014

Institution: ENS-Lyon (France)

Funding: INRIA

Supervision: M. Régnier

Robert Huang

Subject: Repeats in genomic sequences

Date: 01/06/2014 - 25/08/2014

Institution: Berkeley (USA)

Funding: ECOLE POLYTECHNIQUE

Supervision: M. Régnier

Hanlun Jiang

Subject : conformational dynamics of the RNA-induced silencing complex

Date: 01/06/2014 - 25/08/2014

Institution: HKUST (Hong-Kong)

Funding: MRE

Supervision: J. Bernauer

Stéphanie Kamgnia Wonkap

Subject : Extraction de motifs dans les graphes de workflows scientifiques

Date: 01/06/2014 - 30/06/2014

Institution: Univ. Rennes

Funding: INRIA

Supervision: Ch. Froidevaux and S. Cohen-Boulakia

6.4.2. Visits to International Teams

6.4.2.1. Sabbatical programme

Julie Bernauer

Date: Feb 2014 - Jul 2014

Institution: Stanford University (USA)

6.4.2.2. Research stays abroad

Sarah Cohen-Boulakia

Date: Apr 2014

Institution: University of Pennsylvania (USA)

Date: Dec 2014

Institution: Humboldt University of Berlin (Germany)

Yann Ponty

Date: Sep 2013 - Sep 2015

Institution: Simon Fraser University (Canada)

Sabine Peres

Date: Dec 2014

Institution: Friedrich-Schiller-University Jena (Germany)

Alice Heliou

Date: Feb-Apr 2014

Institution: King's College (UK)

Date: December 2014

Institution: Vavilov Institute of General Genetics (Russia)

Amélie Heliou

Date: Mar-May 2014

Institution: Stanford University (USA)

Antoine Soulé

Date: Half-time 2014

Institution: McGill University (Canada)

7. Dissemination

7.1. Promoting Scientific Activities

7.1.1. Scientific events organisation

7.1.1.1. General chair, scientific chair

Yann Ponty

PARC 2014 (PIMS Analytic RNA Combinatorics meeting), Simon Fraser University (Vancouver, Canada)

Christine Froidevaux and Adrien Rougny

Franco-Japanese Workshop 2014: "Logic Based Methods for Systems Biology". This workshop has been organized at PCRI (Orsay, France) from October 6th to October 8th, 2014. There were 14 participants coming from the following institutions: National Institute of Informatics (Tokyo, Japan), Yamanashi University (Yamanashi, Japan), Tokyo Institute of Technology (Tokyo, Japan), Inria AMIB (LRI, France), Institut de Recherche en Communication et Cybernétique de Nantes (Nantes, France), INRA - Centre de Val de Loire (Nouzilly, France) and Laboratoire d'Informatique de Paris 6 (Paris, France). This workshop took place in the context of a National Institute of Informatics (NII) Collaborative Research Project on "Food and Health with Information Technology".

7.1.2. Scientific events selection

7.1.2.1. Chair of conference program committee

Yann Ponty

CMSR 2014 (Computational Methods for Structural RNAs). Satellite event of ECCB'14 (Strasbourg, France).

Loïc Paulevé

SASB 2014 (5th international workshop on Static Analysis and Systems Biology), Munich (Germany)

7.1.2.2. Member of the conference program committee

Sarah Cohen-Boulakia

DILS 2014 (Data integration in the life sciences)

TAPP 2014 (Theory and Practice of Provenance)

SWEET 2014 (Int. sigmod Workshop on scalable workflow enactment engines and technologies)

Alain Denise

ECCB 2014 (European Conference on Computational Biology)

CARI 2014 (African Conference on Research in Computer Science and Applied Mathematics)

CMSR 2014 (Computational Methods for Structural RNAs). Satellite event of ECCB'14 (Strasbourg, France).

Sabine Peres

ECCB 2014 (European Conference on Computational Biology)

Yann Ponty

ECCB 2014 (European Conference on Computational Biology)

ISMB 2014 (International conference on Intelligent Systems for Molecular Biology)

BICOB 2014 (International Conference on Bioinformatics and Computational Biology)

BIOVIS 2014 (Symposium on Biological Data Visualization)

7.1.2.3. Reviewer

Yann Ponty

LATIN 2014 (Latin American Theoretical INformatics Symposium)

Loïc Paulevé

FORMATS 2014 (Formal Modelling and Analysis of Timed Systems)

7.1.3. Journal

7.1.3.1. Member of the editorial board

Sarah Cohen-Boulakia

Journal of Data Semantics (Springer)

Alain Denise

Mathematics of Bio-molecules, speciality of the journal Frontiers in Molecular Biosciences (Ed. Frontiers).

Technique et Science Informatiques (Ed. Hermès)

Yann Ponty

Mathematics of Bio-molecules, speciality of the journal Frontiers in Molecular Biosciences (Ed. Frontiers).

7.1.3.2. Reviewer

The members of the team reviewed numerous papers for numerous journals, including: Bioinformatics, BMC Bioinformatics, RNA, Nucleic Acids Research, Journal of Mathematical Biology, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Journal of Discrete Algorithms, Algorithms for Molecular Biology, PLOS One, Journal of Theoretical Biology, Theoretical Computer Science...

7.1.4. Team seminar

The team seminar, organized by Loïc Paulevé and Sabine Peres, hosted 10 talks delivered by invited speakers from multiple institutions in France and abroad, including:

Alexander Bockmayr (Freie Universität Berlin)

Jan Holub (CTU Prague)

Philippe Icard (BIOTICLA Caen)

Anne Siegel (IRISA)

Denis Thieffry (IBENS)

7.2. Teaching - Supervision - Juries

7.2.1. Teaching

We have and we will go on having trained a group of good multi-disciplinary students both at the Master and PhD level. Being part of this community as a serious training group is obviously an asset. Our project is also very much involved in two major student programs in France: the Master BIBS (Bioinformatique et Biostatistique) at Université Paris-Sud/École Polytechnique and the parcours d'Approfondissement en Bioinformatique at École Polytechnique. We are also involved in a student partnership with McGill University (partenariat France Quebec offering French and Canadian students co-supervised internships (short term -3 to 6 months- or long term -part of the PhD studies-). J.-M. Steyaert is involved in the development of an interdisciplinary cooperation between Polytechnique and AP-HP that will favor interships of Polytechnicians and Masters students in AP-HP operational services.

Ch. Froidevaux is a member of the Scientific Committee of the Computer Science Doctoral School of Paris-Sud University.

Ch. Froidevaux is co-heading the Master (M1 and M2) at the University Paris Sud. At Ecole Polytechnique, J.-M. Steyaert was in charge of M1 and M2 until october 2014. M. Régnier is now in charge. Most team members are teaching in this master.

J.Bernauer was appointed *Chargé d'enseignement* in the Computer Science Department of École Polytechnique (DIX) in 2013.

Master BIBS: J. Bernauer, Informatique théorique et Programmation Python, 20h, M2, Université Paris-Sud, France

Master BIBS: M. Régnier and J.-M. Steyaert, Combinatoire, Algorithmes, Séquences et Modélisation (CASM), 32h, M2, Université Paris-Sud, France

Master : M. Régnier, Basic Algorithms in Computational Biology, 4h, M2, MIPT, Russia.

Cycle Ingénieur Polytechnicien: M. Régnier, Modal Bioinformatique, 8h, 2ème année, École Polytechnique, France

7.2.2. Supervision

PhD

Adrien Guilhot-Gaudeffroy (co-supervised by Ch. Froidevaux, and J. Bernauer, AMIB and J. Azé LIRMM) has defended his PhD thesis (29/09/2014) https://hal.archives-ouvertes.fr/tel-01081605

PhD in progress

Mélanie Boudard, *Game theory and stochastic learning for predicting the three-dimensional structure of large RNA molecules*, Univ. Paris XI, Encadrant(els): D. Barth (Univ. Versailles), J. Cohen (CNRS, LRI) and A. Denise.

Bryan Brancotte, *Ranking biological and biomedical data: algorithms and applications*, Université Paris Sud, 01/10/2012, Encadrants: S. Cohen-Boulakia and A. Denise.

Alice Heliou, *Identification et caractérisation d'ARN circulaires dans des séquences NGS*, Ecole Polytechnique, Encadrants: Mireille Régnier et H. Becker
Amélie Heliou, *Game theory and conformation sampling for multi-scale and multi-body macromolecule docking* , Ecole Polytechnique, Encadrantes: J. Bernauer and J. Cohen
Daria Iakovishina, *A Combinatorial Approach to Assembly Algorithm*, Inria, Encadrante: Mireille Régnier
Vincent Le Gallic, *Design de structures secondaires avec contraintes de séquences : une approche globale fondée sur les langages formels*, Univ. Paris Sud. Encadrants: A. Denise and Y. Ponty
Cécile Pereira, *Nouvelles approches bioinformatiques pour l'étude à grande échelle de l'évolution des activités enzymatiques*, Univ. Paris XI, Encadrants: Olivier Lespinet and Alain Denise
Adrien Rougny, *Reasonings on biological knowledge to build and analyze signalling networks*, Univ. Paris Sud. Encadrante: Ch. Froidevaux
Antoine Soulé, *Evolutionary study of RNA-RNA interactions in yeast, Ecole Polytechnique*, Encadrants: J.-M. Steyaert and J. Waldispuhl (University McGill, Canada).

7.2.3. Juries

Expertise

S. Cohen-Boulakia acted as an external expert in the ERC Consolidator Grant, panel 'Computer Science and Informatics'
Y. Ponty is a member of the 'Comité National' (hiring/evaluation committee) of CNRS in computer science (section 6) and bioinformatics (cid 51); he acted as an external expert for the Emergence program of Ville de Paris, and for the JCJC program of ANR.
M. Régnier was a member of PES/PEDR attribution juries for INRIA and CNRS. She is a member of DIGITEO progam Committee and SDV working group in Saclay area. She acted as an external expert for regional initiatives.
A. Denise was a member of the HCERES evaluation committee of INRA MIAT Unit.

Hiring committees

Maitre de conferences, Paris Sud, 2014 Computer Science department: Sarah Cohen-Boulakia, Sabine Peres, Christine Froidevaux;
Professeur, Bordeaux, 2014, Computer Science department: Alain Denise;
CR2, INRIA-IDF: M. Régnier;
CR2, INRIA-RENNES: M. Régnier;

PhD juries

M. Folschette (Nantes U.) : M. Régnier;
N. Obeid (Toulouse U.): Ch. Froidevaux;
S. Videla (Rennes U.): Ch. Froidevaux;
R. Champeimont (Pierre et Marie Curie U.): A. Denise;
D. Symeonidou (Paris-Sud U.): A. Denise;
A. Jacquot (Paris-Nord U.): A. Denise;

HDR juries

Ch. Sinocquet (Nantes U.): Ch. Froidevaux
M. Smail (Nancy U.): Ch. Froidevaux
P. Amar (Paris-Sud U.) : A. Denise
A. Allauzen (Paris-Sud U.) : A. Denise
F. Tahi (Evry U.) : M. Régnier;
J. Bernauer (Paris-Sud U.) : Ch. Froidevaux, J.-M. Steyaert

7.3. Popularization

Y. Ponty authored the pop. sci. paper *Bio-algorithmique des ARN : petite promenade aux interfaces* [31] for *1024*, the journal of the French society of computer science (SIF).

8. Bibliography

Major publications by the team in recent years

[1] Z. BAO, S. COHEN-BOULAKIA, S. DAVIDSON, P. GIRARD. *PDiffView: Viewing the Difference in Provenance of Workflow Results*, in "PVLDB, Proc. of the 35th Int. Conf. on Very Large Data Bases", 2009, vol. 2, n^{o} 2, pp. 1638-1641

[2] J. BERNAUER, X. HUANG, A. Y. L. SIM, M. LEVITT. *Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation*, in "RNA", June 2011, vol. 17, n^{o} 6, pp. 1066-75 [*DOI :* 10.1261/RNA.2543711], http://hal.inria.fr/inria-00624999

[3] A. DENISE, Y. PONTY, M. TERMIER. *Controlled non uniform random generation of decomposable structures*, in "Journal of Theoretical Computer Science (TCS)", 2010, vol. 411, n^{o} 40-42, pp. 3527-3552 [*DOI :* 10.1016/J.TCS.2010.05.010]

[4] A. LOPES, S. SACQUIN-MORA, V. DIMITROVA, E. LAINE, Y. PONTY, A. CARBONE. *Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information*, in "PLoS Computational Biology", December 2013, vol. 9, n^{o} 12 [*DOI :* 10.1371/JOURNAL.PCBI.1003369], http://hal.inria.fr/hal-00875116

[5] C. SAULE, M. REGNIER, J.-M. STEYAERT, A. DENISE. *Counting RNA pseudoknotted structures*, in "Journal of Computational Biology", October 2011, vol. 18, n^{o} 10, pp. 1339-1351 [*DOI :* 10.1089/CMB.2010.0086], http://hal.inria.fr/inria-00537117

Publications of the year

Articles in International Peer-Reviewed Journals

[6] C. BARTON, A. HELIOU, L. MOUCHARD, S. P. PISSIS. *Linear-time computation of minimal absent words using suffix array*, in "BMC Bioinformatics", 2014, vol. 15, 11 p. [*DOI :* 10.1186/S12859-014-0388-9], https://hal.inria.fr/hal-01110274

[7] O. BERILLO, M. REGNIER, A. IVASHCHENKO. *TmiRUSite and TmiROSite scripts: searching for mRNA fragments with miRNA binding sites with encoded amino acid residues*, in "Bioinformation", July 2014, vol. 10, n^{o} 7, pp. 472-473, https://hal.inria.fr/hal-01071330

[8] S. COHEN-BOULAKIA, J. CHEN, P. MISSIER, C. GOBLE, A. WILLIAMS, C. FROIDEVAUX. *Distilling structure in Taverna scientific workflows: a refactoring approach*, in "BMC Bioinformatics", 2014, vol. 15, n^{o} Suppl 1, S12 p. , https://hal.inria.fr/hal-00926827

[9] A. DENISE, P. RINAUDO. *Optimisation problems for pairwise RNA sequence and structure comparison: a brief survey*, in "Transactions on Computational Collective Intelligence", January 2014, vol. 13, pp. 70-82 [*DOI :* 10.1007/978-3-642-54455-2_3], https://hal.archives-ouvertes.fr/hal-00759573

[10] M. FOLSCHETTE, L. PAULEVÉ, K. INOUE, M. MAGNIN, O. ROUX. *Identification of Biological Regulatory Networks from Process Hitting models*, in "Journal of Theoretical Computer Science (TCS)", February 2015, vol. 568, 39 p. [*DOI :* 10.1016/J.TCS.2014.12.002], https://hal.archives-ouvertes.fr/hal-01094249

[11] R. FONSECA, D. V. PACHOV, J. BERNAUER, H. VAN DEN BEDEM. *Characterizing RNA ensembles from NMR data with kinematic models*, in "Nucleic Acids Research", August 2014, vol. 42, n° 15, pp. 9562-72 [*DOI :* 10.1093/NAR/GKU707], https://hal.inria.fr/hal-01058971

[12] A. GUILHOT-GAUDEFFROY, C. FROIDEVAUX, J. AZÉ, J. BERNAUER. *Protein-RNA Complexes and Efficient Automatic Docking: Expanding RosettaDock Possibilities*, in "PLoS ONE", 2014, vol. 9, n° 9, e108928 [*DOI :* 10.1371/JOURNAL.PONE.0108928], https://hal.inria.fr/hal-01071876

[13] S. LAURENT, B. LUDIVINE, I. PHILIPPE, L. HUBERT, J.-M. STEYAERT. *Metabolic Treatment of Cancer: Intermediate Results of a Prospective Case Series*, in "Anticancer Research", January 2014, https://hal.inria.fr/hal-00933725

[14] C. PEREIRA, A. DENISE, O. LESPINET. *A meta-approach for improving the prediction and the functional annotation of ortholog groups*, in "BMC Genomics", 2014, vol. 15(Suppl 6), S16 p. [*DOI :* 10.1186/1471-2164-15-S6-S16], https://hal.archives-ouvertes.fr/hal-01097742

[15] E. RANAEI-SIADAT, C. MÉRIGOUX, B. SEIJO, L. PONCHON, J.-M. SALIOU, J. BERNAUER, S. SANGLIER-CIANFÉRANI, F. DARDEL, P. VACHETTE, S. NONIN-LECOMTE. *In vivo tmRNA protection by SmpB and pre-ribosome binding conformation in solution*, in "RNA", August 2014, vol. 20, n° 10, pp. 1607-20 [*DOI :* 10.1261/RNA.045674.114], https://hal.inria.fr/hal-01058972

[16] C. SHAO, B. YANG, T. WU, J. HUANG, P. TANG, Y. ZHOU, J. ZHOU, J. QIU, L. JIANG, H. LI, G. CHEN, H. SUN, Y. ZHANG, A. DENISE, D.-E. ZHANG, X.-D. FU. *Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome*, in "Nature Structural and Molecular Biology", 2014, vol. 21, pp. 997-1005 [*DOI :* 10.1038/NSMB.2906], https://hal.archives-ouvertes.fr/hal-01097740

[17] J. STARLINGER, B. BRANCOTTE, S. COHEN-BOULAKIA, U. LESER. *Similarity Search for Scientific Workflows*, in "Proceedings of the VLDB Endowment (PVLDB)", September 2014, 12 p. , https://hal.inria.fr/hal-01066046

[18] T. V. D. TRAN, P. CHASSIGNET, J.-M. STEYAERT. *On permuted super-secondary structures of transmembrane β-barrel proteins*, in "Theoretical Computer Science", 2014 [*DOI :* 10.1016/J.TCS.2013.10.001], https://hal.inria.fr/hal-00869141

Invited Conferences

[19] P. AMAR. *Systèmes oscillants chaotiques en biologie*, in "Ecole de Printemps 2014 de la Société Francophone de Biologie Théorique", St Flour, France, Société Francophone de Biologie Théorique, May 2014, https://hal.inria.fr/hal-01087815

International Conferences with Proceedings

[20] P. AMAR, M. BAILLIEUL, D. BARTH, B. LECUN, F. QUESSETTE, S. VIAL. *Parallel biological in silico simulation*, in "29th Intenational Symposium on Computer and Information Sciences", Krakow, Poland, October 2014, pp. 387-394 [*DOI :* 10.1007/978-3-319-09465-6_40], https://hal.inria.fr/hal-01087811

[21] B. BRANCOTTE, B. RANCE, A. DENISE, S. COHEN-BOULAKIA. *ConQuR-Bio: Consensus Ranking with Query Reformulation for Biological Data*, in "10th International Conference, Data Integration in the Life Sciences", Lisbon, Portugal, July 2014, pp. 128 - 142 [*DOI :* 10.1007/978-3-319-08590-6_13], https://hal.inria.fr/hal-01091053

[22] T. CHATAIN, S. HAAR, L. JEZEQUEL, L. PAULEVÉ, S. SCHWOON. *Characterization of Reachable Attractors Using Petri Net Unfoldings*, in "CMSB 2014", Manchester, United Kingdom, P. MENDES, J. DADA, K. SMALLBONE (editors), LNCS/LNBI, Springer Berlin Heidelberg, November 2014, forthcoming, https://hal.archives-ouvertes.fr/hal-01060450

[23] C. CHAUVE, Y. PONTY, J. P. P. ZANETTI. *Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach*, in "BSB - Brazilian Symposium on Bioinformatics - 2014", Belo Horizonte, Brazil, October 2014, TBA, https://hal.inria.fr/hal-01056140

[24] J. CHEN, S. COHEN-BOULAKIA, C. FROIDEVAUX, C. GOBLE, P. MISSIER, A. WILLIAMS. *DistillFlow: removing redundancy in scientific workflows*, in "SSDBM '14 Proceedings of the 26th International Conference on Scientific and Statistical Database Management", Aalborg, Denmark, June 2014 [*DOI :* 10.1145/2618243.2618287], https://hal.inria.fr/hal-01091033

[25] T. OPITZ, J. AZÉ, S. BRINGAY, C. JOUTARD, C. LAVERGNE, C. MOLLEVI. *Breast cancer and quality of life: medical information extraction from health forums*, in "Medical Informatics Europe", Istanbul, Turkey, August 2014, pp. 1070-1074, https://hal.archives-ouvertes.fr/hal-01061891

[26] M. REGNIER, B. FANG, D. IAKOVISHINA. *Clump Combinatorics, Automata, and Word Asymptotics*, in "ANALCO'14", Portland, United States, M. DRMOTA, M. WARD (editors), SIAM, January 2014, https://hal.inria.fr/hal-00864645

[27] J. STARLINGER, S. COHEN-BOULAKIA, S. KHANNA, S. DAVIDSON, U. LESER. *Layer Decomposition: An Effective Structure-based Approach for Scientific Workflow Similarity*, in "IEEE e-Science conference", Guarujá, Brazil, October 2014, https://hal.inria.fr/hal-01066076

National Conferences with Proceedings

[28] A. GUILHOT-GAUDEFFROY, J. AZÉ, J. BERNAUER, C. FROIDEVAUX. *Apprentissage de fonctions de tri pour la prédiction d'interactions protéine-ARN*, in "Extraction et Gestion des Connaissances", Rennes, France, January 2014, https://hal.inria.fr/hal-01016683

Conferences without Proceedings

[29] Y. YAMAMOTO, A. ROUGNY, H. NABESHIMA, K. INOUE, H. MORIYA, C. FROIDEVAUX, K. IWANUMA. *Completing SBGN-AF Networks by Logic-Based Hypothesis Finding*, in "Formal Methods In Macro-Biology", Nouméa, New Caledonia, Formal Methods in Macro-Biology, Springer, September 2014, vol. 8738, pp. 165-179 [*DOI :* 10.1007/978-3-319-10398-3_14], https://hal.archives-ouvertes.fr/hal-01110133

Scientific Books (or Scientific Book chapters)

[30] L. PAULEVÉ, C. CHANCELLOR, M. FOLSCHETTE, M. MAGNIN, O. ROUX. *Analyzing Large Network Dynamics with Process Hitting*, in "Logical Modeling of Biological Systems", L. F. DEL CERRO, K. INOUE (editors), Wiley, July 2014, pp. 125 - 166, https://hal.archives-ouvertes.fr/hal-01060490

[31] Y. PONTY, F. LECLERC. *Drawing and Editing the Secondary Structure(s) of RNA*, in "RNA Bioinformatics", E. PICARDI (editor), Methods in Molecular Biology, Springer New York, 2015, vol. 1269, pp. 63-100 [*DOI :* 10.1007/978-1-4939-2291-8_5], https://hal.inria.fr/hal-01079893

[32] A. ROUGNY, C. FROIDEVAUX, Y. YAMAMOTO, K. INOUE. *Analyzing SBGN-AF Networks Using Normal Logic Programs*, in "Logical Modeling of Biological Systems", Wiley, July 2014 [*DOI :* 10.1002/9781119005223.CH9], https://hal.archives-ouvertes.fr/hal-01110123

Books or Proceedings Editing

[33] F. JOSSINET, Y. PONTY, J. WALDISPÜHL (editors). *Proceedings of the 1st workshop on Computational Methods for Structural RNAs*, McGill UniversityFrance, September 2014, 76 p. [*DOI :* 10.15455/CMSR.2014.0000], https://hal.inria.fr/hal-01071417

Research Reports

[34] L. PAULEVÉ, M. FOLSCHETTE, M. MAGNIN, O. ROUX. *Analyses statiques de la dynamique des réseaux d'automates indéterministes*, March 2014, https://hal.archives-ouvertes.fr/hal-01070295

Scientific Popularization

[35] Y. PONTY. *Bio-algorithmique des ARN : petite promenade aux interfaces*, in "1024 - Bulletin de la société informatique de France", E. SOPENA (editor), SIF, October 2014, vol. 4, pp. 23–53, https://hal.inria.fr/hal-01077506

Other Publications

[36] J. HALEŠ, J. MAŇUCH, Y. PONTY, L. STACHO. *Combinatorial RNA Design: Designability and Structure-Approximating Algorithm*, 2015, Submitted to CPM'15, https://hal.inria.fr/hal-01115349

[37] S. KAMGNIA WONKAP. *Extraction de motifs dans les graphes de workflows scientifiques*, Laboratoire de Recherche en Informatique [LRI], UMR 8623, Bâtiments 650-660, Université Paris-Sud, 91405 Orsay Cedex, 2014, 38 p. , http://dumas.ccsd.cnrs.fr/dumas-01088813

[38] C. PEREIRA, A. DENISE, O. LESPINET. *A new method for improving the prediction and the functional annotation of ortholog groups*, 2014, Proc. ECCB'14, the 13th European Conference on Computational Biology, https://hal.archives-ouvertes.fr/hal-01098231

[39] A. RAJARAMAN, C. CHAUVE, Y. PONTY. *Assessing the robustness of parsimonious predictions for gene neighborhoods from reconciled phylogenies*, 2015, Submitted to ECCB/ISMB 2015, https://hal.inria.fr/hal-01104587

[40] M. RÉGNIER, E. FURLETOVA, V. YAKOVLEV, M. ROYTBERG. *Analysis of pattern overlaps and exact computation of P-values of pattern occurrences numbers: case of Hidden Markov Models*, December 2014 [*DOI :* 10.1186/s13015-014-0025-1], https://hal.inria.fr/hal-00858701

References in notes

[41] P. AMAR. *Comparative study of some methods for simulation of biochemical reactions*, in "Ecole de Printemps 2012 de la Société Francophone de Biologie Théorique", Saint Flour, France, June 2012, http://hal.inria.fr/hal-00763571

[42] P. AMAR, L. PAULEVÉ. *HSIM: an hybrid stochastic simulation system for systems biology*, in "The Third International Workshop on Static Analysis and Systems Biology (SASB 2012)", Deauville, France, September 2012, http://hal.inria.fr/hal-00758168

[43] Z. ASLAOUI-ERRAFI, S. COHEN-BOULAKIA, C. FROIDEVAUX, P. GLOAGUEN, A. POUPON, A. ROUGNY, M. YAHIAOUI. *Towards a logic-based method to infer provenance-aware molecular networks*, in "Proc. of the 1st ECML/PKDD International workshop on Learning and Discovery in Symbolic Systems Biology (LDSSB)", Bristol, Royaume-Uni, September 2012, pp. 103-110, http://hal.inria.fr/hal-00748041

[44] J. AZÉ, T. BOURQUARD, S. HAMEL, A. POUPON, D. RITCHIE. *Using Kendall-Tau Meta-Bagging to Improve Protein-Protein Docking Predictions*, in "PRIB 2011", DELFT, Pays-Bas, M. LOOG, ET AL. (editors), Marcel Reinders and Dick de Ridder, 2011, pp. 284-295, http://hal.inria.fr/inria-00628038

[45] B. BEAUVOIT, S. COLOMBIÉ, J.-P. MAZAT, C. NAZARET, S. PÉRÈS. *Systematic study of a metabolic network*, in "Advances in Systems and Synthetic Biology", Evry, France, EDP Sciences, March 2014, https://hal.inria.fr/hal-01108859

[46] J. BERNAUER, R. P. BAHADUR, F. RODIER, J. JANIN, A. POUPON. *DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions*, in "Bioinformatics", March 2008, vol. 24, n^{o} 5, pp. 652-8 [*DOI :* 10.1093/BIOINFORMATICS/BTN022], http://hal.inria.fr/inria-00431696

[47] J. BERNAUER, S. C. FLORES, X. HUANG, S. SHIN, R. ZHOU. *Multi-Scale Modelling of Biosystems: from Molecular to Mesocale - Session Introduction*, in "Pacific Symposium on Biocomputing", 2011, pp. 177-80 [*DOI :* 10.1142/9789814335058_0019], http://hal.inria.fr/inria-00542791

[48] V. BOEVA, T. POPOVA, K. BLEAKLEY, P. CHICHE, J. CAPPO, G. SCHLEIERMACHER, I. JANOUEIX-LEROSEY, O. DELATTRE, E. BARILLOT. *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data*, in "Bioinformatics", 2012, vol. 28, n^{o} 3, pp. 423-425, http://dx.doi.org/10.1093/bioinformatics/btr670

[49] T. BOURQUARD, J. BERNAUER, J. AZÉ, A. POUPON. *Comparing Voronoi and Laguerre tessellations in the protein-protein docking context*, in "Sixth annual International Symposium on Voronoi Diagrams", Copenhagen, Danemark, F. Anton and J. Andreas Bærentzen - Technical University of Denmark, June 2009, http://hal.inria.fr/inria-00429618

[50] T. BOURQUARD, J. BERNAUER, J. AZÉ, A. POUPON. *A collaborative filtering approach for protein-protein docking scoring functions*, in "PLoS ONE", 2011, vol. 6, n^{o} 4 [*DOI :* 10.1371/JOURNAL.PONE.0018541], http://hal.inria.fr/inria-00625000

[51] E. A. COUTSIAS, C. SEOK, M. P. JACOBSON, K. A. DILL. *A kinematic view of loop closure*, in "J Comput Chem", Mar 2004, vol. 25, n^{o} 4, pp. 510–528, http://dx.doi.org/10.1002/jcc.10416

[52] K. DARTY, A. DENISE, Y. PONTY. *VARNA: Interactive drawing and editing of the RNA secondary structure*, in "Bioinformatics", August 2009, vol. 25, n^{o} 15, pp. 1974-5 [*DOI :* 10.1093/BIOINFORMATICS/BTP250], http://hal.inria.fr/hal-00432548

[53] A. DENISE, M.-C. GAUDEL, S.-D. GOURAUD, R. LASSAIGNE, J. OUDINET, S. PEYRONNET. *Coverage-biased random exploration of large models and application to testing*, in "Software Tools for Technology Transfer (STTT)", 2012, vol. 14, n^{o} 1, pp. 73-93, http://hal.inria.fr/inria-00560621

[54] A. DENISE, Y. PONTY, M. TERMIER. *Controlled non uniform random generation of decomposable structures*, in "Theoretical Computer Science", 2010, vol. 411, n° 40-42, pp. 3527-3552 [*DOI :* 10.1016/J.TCS.2010.05.010], http://hal.inria.fr/hal-00483581

[55] Y. DING, C. CHAN, C. LAWRENCE. *RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble*, in "RNA", 2005, vol. 11, pp. 1157–1166

[56] S. J. FLEISHMAN, T. A. WHITEHEAD, E.-M. STRAUCH, J. E. CORN, S. QIN, H.-X. ZHOU, J. C. MITCHELL, O. N. A. DEMERDASH, M. TAKEDA-SHITAKA, G. TERASHI, I. H. MOAL, X. LI, P. A. BATES, M. ZACHARIAS, H. PARK, J.-S. KO, H. LEE, C. SEOK, T. BOURQUARD, J. BERNAUER, A. POUPON, J. AZÉ, S. SONER, S. K. OVALI, P. OZBEK, N. B. TAL, T. HALILOGLU, H. HWANG, T. VREVEN, B. G. PIERCE, Z. WENG, L. PÉREZ-CANO, C. PONS, J. FERNÁNDEZ-RECIO, F. JIANG, F. YANG, X. GONG, L. CAO, X. XU, B. LIU, P. WANG, C. LI, C. WANG, C. H. ROBERT, M. GUHAROY, S. LIU, Y. HUANG, L. LI, D. GUO, Y. CHEN, Y. XIAO, N. LONDON, Z. ITZHAKI, O. SCHUELER-FURMAN, Y. INBAR, V. PATAPOV, M. COHEN, G. SCHREIBER, Y. TSUCHIYA, E. KANAMORI, D. M. STANDLEY, H. NAKAMURA, K. KINOSHITA, C. M. DRIGGERS, R. G. HALL, J. L. MORGAN, V. L. HSU, J. ZHAN, Y. YANG, Y. ZHOU, P. L. KASTRITIS, A. M. J. J. BONVIN, W. ZHANG, C. J. CAMACHO, K. P. KILAMBI, A. SIRCAR, J. J. GRAY, M. OHUE, N. UCHIKOGA, Y. MATSUZAKI, T. ISHIDA, Y. AKIYAMA, R. KHASHAN, S. BUSH, D. FOUCHES, A. TROPSHA, J. ESQUIVEL-RODRÍGUEZ, D. KIHARA, P. B. STRANGES, R. JACAK, B. KUHLMAN, S.-Y. HUANG, X. ZOU, S. J. WODAK, J. JANIN, D. BAKER. *Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology*, in "Journal of Molecular Biology", September 2011 [*DOI :* 10.1016/J.JMB.2011.09.031], http://hal.inria.fr/inria-00637848

[57] S. C. FLORES, J. BERNAUER, S. SHIN, R. ZHOU, X. HUANG. *Multiscale modeling of macromolecular biosystems*, in "Briefings in Bioinformatics", July 2012, vol. 13, n° 4, pp. 395-405 [*DOI :* 10.1093/BIB/BBR077], http://hal.inria.fr/hal-00684530

[58] L. JAROSZEWSKI, Z. LI, S. S. KRISHNA, C. BAKOLITSA, J. WOOLEY, A. M. DEACON, I. A. WILSON, A. GODZIK. *Exploration of uncharted regions of the protein universe*, in "PLoS Biol", Sep 2009, vol. 7, n° 9, http://dx.doi.org/10.1371/journal.pbio.1000205

[59] A. LAMIABLE, D. BARTH, A. DENISE, F. QUESSETTE, S. VIAL, E. WESTHOF. *Automated prediction of three-way junction topological families in RNA secondary structures*, in "Computational Biology and Chemistry", January 2012, vol. 37, pp. 1-5 [*DOI :* 10.1016], http://hal.inria.fr/hal-00641738

[60] A. LEVIN, M. LIS, Y. PONTY, C. W. O'DONNELL, S. DEVADAS, B. BERGER, J. WALDISPÜHL. *A global sampling approach to designing and reengineering RNA secondary structures*, in "Nucleic Acids Research", November 2012, vol. 40, n° 20, pp. 10041-52 [*DOI :* 10.1093/NAR/GKS768], http://hal.inria.fr/hal-00733924

[61] S. LORIOT, F. CAZALS, J. BERNAUER. *ESBTL: efficient PDB parser and data structure for the structural and geometric analysis of biological macromolecules*, in "Bioinformatics", April 2010, vol. 26, n° 8, pp. 1127-8 [*DOI :* 10.1093/BIOINFORMATICS/BTQ083], http://hal.inria.fr/inria-00536404

[62] D. J. MANDELL, E. A. COUTSIAS, T. KORTEMME. *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling*, in "Nat Methods", Aug 2009, vol. 6, n° 8, pp. 551–552, http://dx.doi.org/10.1038/nmeth0809-551

[63] D. MANOCHA, Y. ZHU. *Kinematic manipulation of molecular chains subject to rigid constraints*, in "Proc Int Conf Intell Syst Mol Biol", 1994, vol. 2, pp. 285–293

[64] D. MANOCHA, Y. ZHU, W. WRIGHT. *Conformational analysis of molecular chains using nano-kinematics*, in "Comput Appl Biosci", Feb 1995, vol. 11, n^{o} 1, pp. 71–86

[65] J. OUDINET, A. DENISE, M.-C. GAUDEL. *A new dichotomic algorithm for the uniform random generation of words in regular languages (journal version)*, in "Theoretical Computer Science", September 2013, vol. 502, pp. 165-176, http://hal.inria.fr/hal-00716558

[66] M. PARISIEN, F. MAJOR. *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data*, in "Nature", 2008, vol. 452, n^{o} 7183, pp. 51–55

[67] G. PARK, H.-K. HWANG, P. NICODÈME, W. SZPANKOWSKI. *Profile of Tries*, in "SIAM Journal on Computing", 2009, vol. 38, n^{o} 5, pp. 1821-1880 [*DOI :* 10.1137/070685531], http://hal.inria.fr/hal-00781400

[68] Y. PONTY. *Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: The boustrophedon method*, in "Journal of Mathematical Biology", Jan 2008, vol. 56, n^{o} 1-2, pp. 107–127, http://www.lri.fr/~ponty/docs/Ponty-07-JMB-Boustrophedon.pdf

[69] Y. PONTY, M. TERMIER, A. DENISE. *GenRGenS: Software for Generating Random Genomic Sequences and Structures*, in "Bioinformatics", 2006, vol. 22, n^{o} 12, pp. 1534–1535, http://hal.inria.fr/inria-00601060

[70] S. PÉRÈS, M. MARTIN, L. SIMON. *SAT-Based Metabolics Pathways Analysis without Compilation* , in "CMSB 2014", Manchester, United Kingdom, P. M. J. D. K. SMALLBONE (editor), LNCS/LNBI, Springer Berlin Heidelberg, November 2014, https://hal.inria.fr/hal-01108840

[71] V. REINHARZ, Y. PONTY, J. WALDISPÜHL. *Using Structural and Evolutionary Information to Detect and Correct Pyrosequencing Errors in Noncoding RNAs*, in "Journal of Computational Biology", November 2013, vol. 20, n^{o} 11, pp. 905-19, Extended version of RECOMB'13 [*DOI :* 10.1089/CMB.2013.0085], http://hal.inria.fr/hal-00828062

[72] E. SENTER, S. SHEIKH, I. DOTU, Y. PONTY, P. CLOTE. *Using the Fast Fourier Transform to Accelerate the Computational Search for RNA Conformational Switches*, in "PLoS ONE", December 2012, vol. 7, n^{o} 12 [*DOI :* 10.1371/JOURNAL.PONE.0050506], http://hal.inria.fr/hal-00769740

[73] E. SENTER, S. SHEIKH, I. DOTU, Y. PONTY, P. CLOTE. *Using the Fast Fourier Transform to accelerate the computational search for RNA conformational switches (extended abstract)*, in "RECOMB - 17th Annual International Conference on Research in Computational Molecular Biology - 2013", Beijing, Chine, 2013, http://hal.inria.fr/hal-00766780

[74] A. Y. L. SIM, O. SCHWANDER, M. LEVITT, J. BERNAUER. *Evaluating mixture models for building RNA knowledge-based potentials*, in "Journal of Bioinformatics and Computational Biology", April 2012, vol. 10, n^{o} 2, 1241010 [*DOI :* 10.1142/S0219720012410107], http://hal.inria.fr/hal-00757761

[75] T. V. D. TRAN. *Modeling and predicting super-secondary structures of transmembrane beta-barrel proteins*, Ecole Polytechnique X, December 2011, http://hal.inria.fr/tel-00647947

[76] H. VAN DEN BEDEM, I. LOTAN, J. C. LATOMBE, A. M. DEACON. *Real-space protein-model completion: an inverse-kinematics approach*, in "Acta Crystallogr D Biol Crystallogr", Jan 2005, vol. 61, n^{o} Pt 1, pp. 2–13, http://dx.doi.org/10.1107/S0907444904025697

[77] J. WALDISPÜHL, Y. PONTY. *An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure*, in "Journal of Computational Biology", November 2011, vol. 18, n^{o} 11, pp. 1465-79 [*DOI :* 10.1089/CMB.2011.0181], http://hal.inria.fr/hal-00681928