IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Ecole Polytechnique

Activity Report 2012

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, Research Director (DR) Inria, HdR]
Julie Bernauer [Research Associate (CR) Inria]
Yann Ponty [Research Associate (CR) CNRS]

Faculty Members

Patrick Amar [Université Paris-Sud]
Jérôme Azé [Université Paris-Sud, HdR]
Sarah Cohen-Boulakia [Université Paris-Sud]
Alain Denise [Université Paris-Sud, HdR]
Christine Froidevaux [Université Paris-Sud, HdR]
Sabine Peres [Université Paris-Sud]
Jean-Marc Steyaert [Ecole Polytechnique, HdR]
Florence d'Alché-Buc [on leave from Université Evry, 50%, HdR]

PhD Students

Bryan Brancotte [Université Paris -Sud XI, from 01/10/12]
Jiuqiang Chen [Université Paris-Sud]
Adrien Guilhot-Gaudeffroy [Université Paris -Sud XI, from 01/10/11]
Daria Iakovishina [Ecole Polytechnique]
Feng Lou [Université Paris-Sud,until 31/01/2012]
Philippe Rinaudo [Université Paris-Sud]
Cécile Pereira [Université Paris-Sud]
Cong Zeng [Université Paris-Sud]
Bo Yang [Université Paris-Sud and Wuhan University]
Soulé Antoine [Ecole Polytechnique and University McGill]

Post-Doctoral Fellows

Loic Paulevé [Ecole Polytechnique until 01/07/2012]
Rasmus Fonseca [Inria from 01/12/2012]

Administrative Assistant

Evelyne Rayssac [Secretary (SAR) Inria]

2. Overall Objectives

2.1. Overall Objectives

This project in bioinformatics is mainly concerned with the molecular levels of organization in the cell, dealing principally with RNAs and proteins; we currently concentrate our efforts on structure, interactions, evolution and annotation and aim at a contribution to protein and RNA engineering. 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. On the other hand, we apply these computational approaches to several particular problems arising in fundamental molecular

biology. These problems, described below, raise different computer science issues. To tackle them, the project members rely on a common methodology for which our group has a significant experience. 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.

We investigate the relations between nucleotide sequences, 3D structures and, finally, biochemichal function. All protein functions and many RNA functions are intimately related to the three-dimensional molecular structure. Therefore, we view structure prediction and sequence analysis as an integral part of gene annotation that we study simultaneously and that we plan to pursue on a RNAomic and proteomic scale. Our starting point is the sequence either *ab initio* or with knowledge such as a 3D structural template or ChIP-Chip experiments. We are interested in deciphering the information organization in DNA sequences and identifying the role played by gene products: proteins and RNA, including noncoding RNAs. A common toolkit of computational methods is developed, that relies notably on combinatorial algorithms, mathematical analysis of algorithms and data mining. One goal is to provide softwares or platform elements to predict either structures or structural and functional annotation. For instance, a by-product of 3D structure prediction for protein and RNA engineering is to allow to propose sequences with admissible structures. Statistical softwares for structural annotation are included in annotation tools developped by partners, notably our associate team MIGEC.

Our work is organized along two main axes. The first one is structure prediction, comparison and design engineering. The relation between nucleotide sequence and 3D macromolecular structure, and the relation between 3D structure and biochemical function are possibly the two foremost problems in molecular biology. There are considerable experimental difficulties in determining 3D structures to a high precision. Therefore, there is a crucial need for efficient computational methods for structure prediction, functional assignment and molecular engineering. A focus is given on both protein and RNA structures.

The second axis is structural and functional annotation, a special attention being paid to regulation. Structural annotation deals with the identification of genomic elements, e.g. genes, coding regions, non coding regions, regulatory motifs. Functional annotation consists in characterizing their function, e.g. attaching biological information to these genomic elements. Namely, it provides biochemical function, biological function, regulation and interactions involved and expression conditions. High-throughput technologies make automated annotation crucial. There is a need for relevant computational annotation methods that take into account as many characteristics of gene products as possible -intrinsic properties, evolutionary changes or relationships- and that can estimate the reliability of their own results.

3. Research Program

3.1. RNA and protein structures

3.1.1. RNA

Participants: Julie Bernauer, Alain Denise, Rasmus Fonseca, Feng Lou, Yann Ponty, Mireille Régnier, Philippe Rinaudo, Jean-Marc Steyaert.

Common activity with P. Clote (Boston College and Digiteo).

3.1.1.1. From RNA structure to function

We are currently developing a combinatorial approach, based on random generation, to design small and structured RNAs. An application of such a methodology to the Gag-Pol HIV-1 frameshifting site will be carried out with our collaborators at IGM. We hope that, upon capturing the hybridization energy at the design stage, one will be able to gain control over the rate of frameshift and consequently fine-tune the expression of *Gag/Pol*. Our goal is to build these RNA sequences such that their hybridization with existing mRNAs will be favorable to independent folding, and will therefore affect the stability of some secondary structures involved in recoding events. Moreover it has been observed, mainly on bacteria, that some mRNA sequences may adopt alternate folds. Such events are called a conformational switch, or riboswitch. A common feature of

recoding events and riboswitches is that some structural element on the mRNA initiates and unusual action of the ribosome, or allows for an alternate fold under some environmental conditions. One challenge is to predict genes that might be subject to riboswitches.

3.1.1.2. Beyond secondary structure

One of our major challenges is to go beyond secondary structure. Over the past decade, few attempts have been made to predict the 3D structure of RNA from sequence only. So far, few groups have taken this leap. Despite the promises shown by their preliminary results, these approaches currently suffer to a limiting scale due to either their high algorithmic complexity or their difficult automation. Using our expertise in algorithmics and modeling, we plan to design original methods, notably within the AMIS-ARN project (ANR BLANC 2008-2012) in collaboration with PRISM at Versailles University and E.Westhof's group at Strasbourg.

1. *Ab initio* modeling: Starting from the predicted RNA secondary structure, we aim to detect *local structural motifs* in it, giving local 3D conformations. We use the resulting partial structure as a flexible scaffold for a multi-scale reconstruction, notably using game theory. We believe the latter paradigm offers a more realistic view of biological processes than global optimization, used by our competitors, and constitutes a real originality of our project.
2. Comparative modeling: we investigate new algorithms for predicting 3D structures by a comparative approach. This involves comparing multiple RNA sequences and structures at a large scale, that is not possible with current algorithms. Successful methods must rely both on new graph algorithms and on biological expertise on sequence-structure relations in RNA molecules.

3.1.1.3. RNA 3D structure evaluation

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. Their function is mainly determined by the structure those molecules adopt as protein and nucleic acids differ from polypeptides and polynucleotides by their spatial organization. This is specially challenging for RNA where structure flexibility is key.

To address those issues, one has to explore the biologically possible spatial configurations of a macromolecule. The two most common techniques currently used in computational structural biology are Molecular Dynamics (MD) and Monte Carlo techniques (MC). Those techniques require the evaluation of a potential or force-field, which for computational biology are often empirical. They mainly consist of a summation of bonded forces associated with chemical bonds, bond angles, and bond dihedrals, and non-bonded forces associated with van der Waals forces and electrostatic charges. Even if there exists implicit solvent models, they are yet not very well performing and still require a lot of computation time.

Our goal, in collaboration with the Levitt lab at Stanford University and H. van den Bedem at the Stanford Synchrotron Radiation Laboratory (Associate Team ITSNAP http://pages.saclay.inria.fr/julie.bernauer/EA_ITSNAP/) is to develop knowledge-based (KB) potentials, based on measurements on known RNA 3D structures and provide sampling for experimental structure fitting and docking conformation generation. KB potential are quick to evaluate during a simulation and can be used without having to explicitly address the solvent problem. They can be developed at various levels of representation: -atom, base, nucleotide, domain- and could allow the modelling of a wide size range: from a hairpin to the whole ribosome. We also intend to combine these knowledge-based potentials with other potentials (hybrid modelling) and template-based techniques, allowing accurate modelling and dynamics study of very large RNA molecules. Such studies are still a challenge. We will also study conformations for experimental data fitting by extension the innovative, robotics-inspired Kino-Geometric Sampler conformational search algorithm for proteins to nucleic acids and to include experimental data. KGS models a protein as a kinematic linkage and additionally considers hydrogen bonds that "close" kinematic cycles. In closed kinematic cycles rotatable bonds can no longer be deformed independently without breaking closure. KGS preserves all kinematic cycles, and thus hydrogen bonds, by sampling in a subspace of conformational space defined by all closure constraints. KGS exhibits a singularly large search radius and optimally reduces the number of free parameters. These unique features enable flexible 'docking' of atomic models in the data while moderating the risk of overfitting at low resolution. The KGS procedure can also accommodate knowledge-based potentials to improve evaluation of putative conformations and their interactions (see below).

3.1.2. PROTEINS

Participants: Jérôme Azé, Julie Bernauer, Adrien Guilhot-Gaudeffroy, Jean-Marc Steyaert.

3.1.2.1. Docking and evolutionary algorithms

As mentioned above, the function of many proteins depends on their interaction with one or many partners. Docking is the study of how molecules interact. Despite the improvements due to structural genomics initiatives, the experimental solving of complex structures remains a difficult problem. The prediction of complexes, *docking*, proceeds in two steps: a configuration generation phase or *exploration* and an evaluation phase or *scoring*. As the verification of a predicted conformation is time consuming and very expensive, it is a real challenge to reduce the time dedicated to the analysis of complexes by the biologists. Various algorithms and techniques have been used to perform exploration and scoring [49]. The recent rounds of the CAPRI challenge show that real progress has been made using new techniques [46], [3]. Our group has strong experience in cutting edge geometric modelling and scoring techniques using machine learning strategies for protein-protein complexes. In a collaboration with A. Poupon, INRA-Tours, a method that sorts the various potential conformations by decreasing probability of being real complexes has been developed. It relies on a ranking function that is learnt by an evolutionary algorithm. The learning data are given by a geometric modelling of each conformation obtained by the docking algorithm proposed by the biologists. Objective tests are needed for such predictive approaches. The *Critical Assessment of Predicted Interaction*, CAPRI, a community wide experiment modelled after CASP was set up in 2001 to achieve this goal (http://www.ebi.ac.uk/msd-srv/capri/). First results achieved for CAPRI'02 suggested that it is possible to find good conformations by using geometric information for complexes. This approach has been followed (see section New results). As this new algorithm will produce a huge amount of conformations, an adaptation of the ranking function learning step is needed to handle them. In the near future, we intend to extend our approach to protein-RNA complexes.

Such as in the protein case, the function of RNA molecules also depends on their interaction with one or many partners. Upon interaction, RNA molecules often undergo large conformation changes. Understanding how these molecules interact with proteins would allow better targeting for therapeutic studies. The CAPRI (Critical Assessment of PRediction of Interactions) challenge1 has shown that classical docking procedures largely fail when large conformation change occurs and when RNA is involved. This is especially true for RNA molecules, whose large-scale dynamics remain often unknown. Modeling RNA conformational changes is made hard by the inherent flexible nature of their structure but also by the electrostatics involved. These are hard to model and often lead to computationally expensive simulations. Even if for small RNA molecules, molecular dynamics can be used, such simulations are hard to extend to larger molecules and protein-RNA complexes.

For many diseases, such as cancer and HIV, microRNA molecules play a very important role regulating gene expression by guiding the RISC. Some miRNAs have been shown to suppress tumors and are thus ideal candidates for the development of therapeutic agents. Even if various computational techniques have been developed to predict miRNA targets, none of these consider the structural aspects of the interactions between components of the RISC and miRNA. We aim to target these problems, in collaboration with the Huang lab at HKUST (PHC Procore)

The combination of Voronoi models at a coarse-grained level and powerful machine learning techniques allows the accurate scoring of protein-protein complexes [12]. Our actual machine learning approach for proteins is a combination of several machine learning approaches (evolutionary algorithm, decision trees, decision rules,...). By adapting these approaches to protein-RNA, we would have a fast and efficient technique for scoring large protein-RNA complexes where conformational changes are involved.

Working with RNA instead of protein introduce many major differences in the machine learning approaches. RNA conformations are often smaller than protein conformations, which has an impact on the values of the descriptors used to describe objects. Due to the size differences between RNA and protein, it is often more difficult to generate (during the modeling stage) conformations closed to the biological solution (near native solution). The machine learning algorithms therefore need to take into account all theses specificities to be able to learn good predictive models from data that are not very close to the real solution.

The acquired knowledge on RNA flexibility, dynamics and the importance of the sequence will be a strong advance in the modeling of protein-RNA interactions we are working on. Il will help the development of scoring functions based on Voronoi models for RNA and provide us with the level of flexibility needed in complex conformational search. We also intend to develop hybrid KB potentials for complexes from hybrid RNA KB data. These could be incorporated in leading-edge flexible docking modeling software such as Rosetta.

3.2. Annotations

3.2.1. Word counting

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

We aim at enumerating or generating sequences or structures that are *admissible* in the sense that they are likely to possess some given biological property. Team members have a common expertise in enumeration and random generation of combinatorial structures. They have developped computational tools for probability distributions on combinatorial objects, using in particular generating functions and analytic combinatorics. Admissibility criteria can be mainly statistic; they can also rely on the optimisation of some biological parameter, such as an energy function.

The ability to distinguish a significant event from statistical noise is a crucial need in bioinformatics. In a first step, one defines a suitable probabilistic model (null model) that takes into account the relevant biological properties on the structures of interest. A second step is to develop accurate criteria for assessing (or not) their exceptionality. An event observed in biological sequences, is considered as exceptional, and therefore biologically significant, if the probability that it occurs is very small in the null model. Our approach to compute such a probability consists in an enumeration of good structures or combinatorial objects. Thirdly, it is necessary to design and implement efficient algorithms to compute these formulae or to generate random data sets. Typical examples that motivate research on words and motifs counting are *Transcription Factor Binding Sites*, TFBSs, consensus models of recoding events and some RNA structural motifs. When relevant motifs do not resort to regular languages, one may still take advantage of combinatorial properties to define functions whose study are amenable to our algebraic tools. One may cite secondary structures and recoding events.

Fast development of high throughput technologies has generated a new challenge for computational biology. The main bottlenecks in applications are the computational analysis of experimental data.

As a first example, numerous new assembling algorithms have recently appeared. Still, the comparison of the results arising from these different algorithms led to significant differences for a given genome assembly. Clearly, strong constraints from the underlying technologies, leading to different data (size, confidence,...) are one origin of the problems and a deeper interpretation is needed, in order to improve algorithms and confidence in the results. One objective is to develop a model of errors, including a statistical model, that takes into account the quality of data for the different technologies, and their volume. This is the subject of an international collaboration with V. Makeev's lab (IoGene, Moscow) and MAGNOME project-team. Second, Next Generation Sequencing open the way to the study of structuralvariants in the genome, as recently described in [44]. Defining a probabilistic model that takes into account main dependencies -such as the GC content- is a task o D. Iakovishina's thesis, in a starting 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. Stochastic context-free grammars (SCFG's) have long been used to model both structural and statistical properties of genomic sequences, particularly for predicting the structure of sequences or for searching for motifs. They can also be used to generate random sequences. However, they do not allow the user to fix the length of these sequences. We developed algorithms for random structures generation that respect a given probability distribution on their components. Our approach is based on the concept of *weighted* combinatorial classes, in combination with the so-named *recursive* method for generating combinatorial structures. To that purpose, one first translates the (biological) structures into combinatorial classes, using the *symbolic method*, an algebraic framework developed by Flajolet *et al*. Adding weights to the atoms allows one to bias the probabilities towards the desired distribution. The main issue is to develop efficient algorithms for finding the suitable weights. An implementation was given in the `GenRGenS` software http://www.lri.fr/~genrgens/, and a generic optimizer that automatically derives suitable parameters for a given grammar, is currently being developped.

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [45]: 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 [48], [4], 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 [47].

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 [10], [21].

3.2.3. Programmed -1 ribosomal frameshifting

Participants: Patrick Amar, Jérôme Azé, Alain Denise, Christine Froidevaux, Yann Ponty, Cong Zeng.

During protein synthesis, the ribosome decodes the mRNA by assigning a specific amino acid to each codon or nucleotide triplet. Throughout this process the ribosome moves along the mRNA molecule three nucleotides at a time. However, encounters of specific signals found in mRNA from many viruses lead the ribosome to shift one nucleotide backward thus changing its reading frame. We aim at developing a new computational approach that is able to detect these specific signals in genomic databases in order to better understand the molecular choreography leading to the ribosomal frameshifting, which ultimately will help to rationally design new antiviral drugs. As candidates sequences are expected to be numerous, we aim at developing a ranking method to identify the most relevant sequences. Biological testing of these most promising identified candidates by our collaborators from IGM will help us to refine our computational method.

3.2.4. Knowledge extraction

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

Our main goal is to design semi-automatic methods for annotation. A possible approach is to focus on the way we could discover relevant motifs in order to make more precise links between function and motifs sequence. For instance, a commonly accepted hypothesis is that function depends on the order of the motifs present in a genomic sequence. Likewise we must be able to evaluate the quality of the annotation obtained. This necessitates giving an estimate of the reliability of the results. This may use combinatorial tools described above. It includes a rigorous statement of the validity domain of algorithms and knowledge of the results provenance. We are interested in provenance resulting from workflow management systems that are important in scientific applications for managing large-scale experiments and can be useful to calculate functional annotations. A given workflow may be executed many times, generating huge amounts of information about data produced and consumed. Given the growing availability of this information, there is an increasing interest in mining it to understand the difference in results produced by different executions.

3.2.5. Systems Biology

Participants: Patrick Amar, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Loic Paulevé, Sabine Peres, Mireille Régnier, Jean-Marc Steyaert.

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).

3.2.5.1. Simulations and behavior analysis for metabolism modeling

A great number of methods have been proposed for the study of the behavior of large biological systems. Two methods have been developed and are in use in the team, depending on the specific problems under study : the first one is based on a discrete and direct simulation of the various interactions between the reactants, while the second one deals with an abstract representation by means of differential equations from which we extract various types of features pertaining to the system.

We investigate on the computational modelling step of the cycle by developing a computer simulation system, HSIM, that mimics the interactions of biomolecules in an environment modelling the membranes and compartments found in real cells. In collaboration with biologists from the AMMIS lab. at Rouen we have used HSIM to show the properties of grouping the enzymes of the phosphotransferase system and the glycolytic pathway into metabolons in *E. coli*. In another collaboration with the SYSDIAG Lab (UMR 3145) at Montpellier, we participate at the CompuBioTic project. This is a Synthetic Biology project in the field of medical diagnosis: its goal is to design a small vesicle containing specific proteins 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 human pathologies. We used HSIM to help the design and to test qualitatively and quantitatively this *"biological computer"* before *in vitro*.

Given the set of biochemical reactions which 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. We have developed a software `Mpas` (Metabolic Pathway Analyser Software) that renders the translation in terms of a systems of o.d.e's automatic; then the simulations are also made easy and almost automatic. 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 particular we have been able to prove mathematically the decisive influence of the enzyme PEMT on the Choline/Ethylamine cycles (M. Behzadi's thesis, defended in 2011).

3.2.5.2. Comparison of Metabolic Networks

In the context of a national project, 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 will be studied to find new enzymes allowing the transformation of biomass. The KEGG database contains pathways related to fungi and other species. By analysing these known pathways with rules mining approaches, we would be able to predict new enzymes activities.

3.2.5.3. Signalling networks

AMIB and INRA-BIOS (A. Poupon, Tours) are partners in a two years project ASAM (2011-2012). This project aims to help the understanding of signalling pathways involving G protein-coupled receptors (GPCR) which 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. The aim of ASAM is thus to provide means to semi-automatically construct signalling networks of GPCRS.

4. Software and Platforms

4.1. Varna

Participants: Yann Ponty [correspondant], Alain Denise.

VARNA is a tool for the automated drawing, visualization and annotation of the secondary structure of RNA, designed as a companion software for web servers and databases. VARNA implements four drawing algorithms, supports input/output using the classic formats *dbn, ct, bpseq* and *RNAML* and exports the drawing, either as a bitmap (*JPEG, PNG*) or as a vector picture (*SVG, EPS* and *XFIG*). It also allows manual modification and structural annotation of the resulting drawings using either an interactive *point and click* approach, within a web server or through command-line arguments. VARNA is a free software distributed under the terms of the GPLv3.0 license and available at http://varna.lri.fr.

VARNA is currently used by RNA scientists (Cited by 92 research articles since its presentation in Fall of 2009, according to Google scholar), web servers such as the BOULDEALE (http://www.microbio.me/boulderale/), TFOLD (http://tfold.ibisc.univ-evry.fr/TFold/), CYLOFOLD (http://cylofold.abcc.ncifcrf.gov/) webservers, and by databases such as IRESITE (http://iresite.org/), SRNATARBASE (http://ccb.bmi.ac.cn/srnatarbase/)and RFAM (http://rfam.sanger.ac.uk/), the main source of sequence/structure data for RNA scientist, to display secondary structures. It is also used as an integrated component within JALVIEW, arguably one of the leading sequence alignment editor (http://www.jalview.org/), and Y. Ponty co-supervised with Jim Procter (University of Dundee, Jalview Project Leader) two internships (including a Google Summer of Code) in the summer of 2012 to further the interactions between the two software.

4.2. SPFlow

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

SPFLOW is a scientific workflow rewriting tool. SPFlow aims at transforming complex workflow structures (non series-parallel structures) into provenance-equivalent simple workflow structures (series-parallel structures). SPFlow takes as an input a file representing one scientific workflow from Taverna and produces another file in which the structure of the original workflow is made series-parallel while ensuring that both workflows have the same provenance (more information available at [32], [39]). The tool is freely available at http://www.lri.fr/~chenj/SPFlow.

4.3. GeneValorization

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

High-throughput technologies provide fundamental information concerning thousands of genes. Many of the current research laboratories daily use one or more of these technologies and end-up with lists of genes. Assessing the originality of the results obtained includes being aware of the number of publications available concerning individual or multiple genes and accessing information about these publications. Faced with the exponential growth of publications available and number of genes involved in a study, this task is becoming particularly difficult to achieve. We introduce GENEVALORIZATION, a web-based tool which gives a clear and handful overview of the bibliography available corresponding to the user input formed by (i) a gene list (expressed by gene names or ids from ENTREZGENE) and (ii) a context of study (expressed by keywords). From this input, GENEVALORIZATION provides a matrix containing the number of publications with co-occurrences of gene names and keywords. Graphics are automatically generated to assess the relative importance of genes within various contexts. Links to publications and other databases offering information on genes and keywords are also available. To illustrate how helpful GENEVALORIZATION is, we have considered the gene list of the OncotypeDX prognostic marker test. it is available at http://bioguide-project.net/gv.

4.4. HSIM

Participant: Patrick Amar [correspondant].

HSIM 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 which 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.

4.5. 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 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 [14].

5. New Results

5.1. RNA structures

5.1.1. RNA structure alignment

It is widely accepted that, for a large number of RNA families, the structure is more conserved than the sequence. Therefore, any reasonable notion of homology should consider the similarity in the secondary structure, i.e. how well the base-pairing positions in two structures can be put in correspondence, or aligned. In collaboration with a significant part of the French bioinformatics community, an assessment of the quality of existing algorithms for the problem was proposed [6]. Furthermore, a review of the state-of-the-art in RNA comparison algorithms is to be published [11], and a chapter in a forthcoming book on RNA computational biology was written in collaboration with Robert Giegerich (University Bielefeld) during his stay.

Most existing alignment tools rely on the assumption that the RNA structure is free of pseudoknots, i.e. free of crossing interactions. This condition naturally arises from the intractability of the unconstrained version of the problem. In a joint work, A. Denise, Ph. Rinaudo and Y. Ponty worked around this issue by proposing a parameterized complexity algorithmic solution for the unconstrained version of the problem. One of the key feature of this algorithm is that, although exponential in the worst-case scenario, it naturally adapts its complexity to the level of intricacy of the aligned structures, and remains polynomial for large classes of pseudoknots. Preliminary results of this work were presented at the WABI'12 conference [35].

5.1.2. Energy-weighted RNA algorithmics

We complemented previous studies led within AMIB on RNA structures with restricted classes of pseudoknots by showing, in a collaboration with Rolf Backofen (Freiburg University), that the computational hardness of RNA folding with general pseudoknots is extremely robust to the choice of a precise energy model. It was shown that the problem is completely unapproximable when expressive – yet realistic – energy models are taken into consideration. These results were presented at CPM'12 [37] (Helsinki, Finland).

Moreover, using an interpolation technique introduced at the RECOMB'11 conference, we were able to improve both the sequential and parallel complexities of the `RNAbor` algorithm developed within P. Clote's lab. The resulting algorithm and its application to the detection of conformational switches in sequence lengths that were previously unreachable by the algorithm, are described in a manuscript accepted in *Plos One*.

5.1.3. RNA knowledge-based potentials and 3D studies

The building of an RNA potential proved much harder and interesting than we initially expected. A non-redundant dataset had first to be extracted from the literature as the available dataset were not suitable for our study even the very recent ones. From the collected distance data, the building of a knowledge-based potential was usually done using histograms; and the histogram interval size and data fitting was an issue. In our 2012 study, we showed that the best solution to build potentials with no interval issue is by using Dirichlet Process Mixture Models (DPMs) [24]. We also benefited of the group experience in modeling the dynamics of RNA and normal-mode experiments to obtain two good decoy sets which complemented the well-known Farna study. We also showed that in many case our high-resolution predictions were better than the Farna/Rosetta standard.

5.1.4. RNA 3D structure prediction

In collaboration with PRISM at Versailles and Westhof's group at Strasbourg, we addressed the problem of ab initio prediction of RNA three-dimensional structure. We developed an algorithm for automatically predicting the topological family of any RNA three-way junction, an thus its coarse-grained local geometry, given only the information from the secondary structure: the sequence and the Watson–Crick pairings. Additionally, we showed that the results are noticeably improved if homology information is used [14]. The resulting software, `Cartaj`, is available online and downloadable at http://cartaj.lri.fr. Then we investigated a new approach for the global prediction of the coarse-grain 3D structure of RNA molecules. We model a molecule as being made of helices and junctions. Using our results above, we are able to classify junctions into topological families that determine their preferred 3D shapes. All the parts of the molecule are then allowed to establish long-distance contacts that induce a threedimensional folding of the molecule. An algorithm relying on game-theory was proposed to discover such long-distance contacts that allow the molecule to reach a Nash equilibrium. As reported by our experiments, this approach allows one to predict the global shape of large molecules of several hundreds of nucleotides that are out of reach of the state-of-the-art methods [15].

A graph-theoretic approach has been successfully used for classification and structure prediction of transmembrane beta-barrel proteins[23], [25].

5.2. Proteins structures and interactions

5.2.1. Protein-protein interaction

Adrien Guilhot, PhD candidate in our project worked on a modified scoring function for the Rosetta software suite. After an extensive conformation generation for the two recently published benchmarks, we now have a model for protein-RNA semi-flexible docking which is currently being tested.

The prediction of the network of protein-protein interactions (PPI) of an organism is crucial for the understanding of biological processes and for the development of new drugs. Machine learning methods have been successfully applied to the prediction of PPI in yeast by the integration of multiple direct and indirect biological data sources. However, experimental data are not available for most organisms. We propose in [9] an ensemble machine learning approach for the prediction of PPI that depends solely on features independent from experimental data. New estimators of the coevolution between proteins have been developed and combined them in an ensemble learning procedure.

This method has been applied to a dataset of known co-complexed proteins in *Escherichia coli* and compared it to previously published methods. Our method allows prediction of PPI with an unprecedented precision of 95.5% for the first 200 sorted pairs of proteins compared to 28.5% on the same dataset with the previous best method.

A close inspection of the best predicted pairs allowed us to detect new or recently discovered interactions between chemotactic components, the flagellar apparatus and RNA polymerase complexes in *E. coli*.

5.3. Combinatorics and Annotation

5.3.1. Word counting and random generation

A long-term research on word enumeration has been realized by the team, in order to calculate a statistical significance for a pattern occurrence according to a given background model. As a part of E. Furletova's thesis, defended in February 2012, co-advised by M. Roytberg (IMPB, Puschino, Russia) and M. Régnier, an extension to Hidden Markov Models, `SufPref`, has been proposed. It relies on a new concept of overlap graphs that efficiently overcomes the main difficulty - overlapping occurrences - in probabilities computation. An implementation is available at http://server2.lpm.org.ru/bio/online/sf/. This algorithm provides a significant space improvement over a previous algorithm, `AhoPro` developed with our former associate team MIGEC. Word statistics were used to identify mRNA targets for miRNAs involved in carcinogenesis [13].

Large deviation results have been derived in [41] that take advantage of general combinatorial properties of words. First, an approximation is derived for the double strands counting problem that refers to a counting of a given pattern in a set of sequences that arise from both strands of the genome. Here dependencies between a sequence and its complement plays a fundamental role. Second, sets of small sequences, with non-identical distributions, are addressed. Possible applications are the search of cis-acting elements in regulatory sequences that may be known, for example from ChIP-chip or ChipSeq experiments, as being under a similar regulatory control.

In [21], we developed a new algorithm for generating uniformly at random words of any regular language L. When using floating point arithmetics, its bit-complexity is $O(q \log n)$ in space and $O(qn \log n)$ in time, where n stands for the length of the word, and q stands for the number of states of a finite deterministic automaton of L. We implemented the algorithm and compared its behavior to the state-of-the-art algorithms, on a set of large automata from the VLTS benchmark suite. Both theoretical and experimental results show that our algorithm offers an excellent compromise in terms of space and time requirements, compared to the known best alternatives. In particular, it is the only method that can generate long paths in large automata. Moreover, in [10], in collaboration with the Fortesse group at LRI, we presented several randomised algorithms for generating paths in large models according to a given coverage criterion. This work opens new perspectives for future studies of statistical testing and model checking, mainly to fight the combinatorial explosion problem.

5.3.2. Analysis and design of weighted combinatorial models

Weighted context-free grammars are natural – yet powerful – random models for biological sequence and structures. We furthered our developments on these objects, and applied them to the study of the Boltzmann ensemble of low-energy in RNA.

In collaboration with P. Clote (Boston College), we used such analytic combinatorics to establish that the average geometric distance between the terminal ends of an RNA sequence, once folded, is asymptotically constant [8].

Furthermore, in collaboration with C. Banderier, O. Bodini and H. Tafat (LIPN), we constructively showed that any predefined distribution of pattern could be attained by a (possibly ambiguous) regular expressions. We also designed a dynamic-programming algorithm to automatically build such models, adopting a segmentation approach based on a parsimony principle. This work was presented at the ANALCO'12 conference [30].

Finally, we continued with D. Gardy and J. Du Boisberranger (PRISM, Université de Versailles-St Quentin) a joint study of collisions in weighted random generation. Indeed, while performing a random generation within large collections of weighted objects, the probability of any sample can be exactly and efficiently computed. Therefore, any redundancy in the sampled set is uninformative (contrasting with situations where the probability is also estimated by the sampling procedure). Following previous results presented at GASCOM'10 (Montreal), we presented at the AOFA'12 (Montreal, Canada) conference [33], a new close formula for the waiting-time of the coupon collector problem, i.e. the average number of words that one must draw to obtain the full collection. The framework defined here has direct applications in the context of RNA : approaches based on sampling are preferred to deterministic optimizations, and algorithmic efficiency of the methods can be critically affected by the redundancy of sampled sets. .

5.3.3. Scientific Workflows

Several Scientific workflow systems have been designed to support users in the tasks of designing, managing, monitoring, and executing in-silico experiments. Such systems are now equipped of provenance modules able to collect data produced and consumed during workflow runs to enhance reproducibility. In this context, we have worked in two directions. First, we have worked on the problem of reuse between scientific workflows. In particular, we have identified the presence of common or similar (sub-)workflows and workflow elements, and have deeply studied, for the first time in the literature, the problem of cross-author reuse [38].

Second, we have worked on studying the structure of scientific workflows. More precisely, we have focused on the series-parallel graph structures. Designing sub-workflows, querying or monitoring workflows leads to perform graph sub-isomorphism. This problem is NP-complete when general DAGs are considered but can be solved in polynomial time when graphs restricted to SP graphs are considered. We have designed and implemented the SPFlow algorithm that rewrites any workflow into an SP workflow while ensuring that the provenance of the rewritten workflow is the same as the original [32], [39].

We are currently working on identifying the reasons why some scientific workflows have a non SP structure. Our long-term goal is to design a *distilling procedure* for scientific workflows offering users the ability of naturally designing workflows having a structure close to SP structures. This work is done in close collaboration with the University of Manchester [31].

5.4. Systems Biology

5.4.1. Reasoning on knowledge to build signaling networks:

We have introduced a logic-based method to infer molecular networks and show how it allows inferring signalling 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 [29].

5.4.2. Metabolic pathways

The topological analyse of metabolic networks is a first step to understand their behaviours and is described in term of fluxes analyses. We work on the elaboration of a stoichiometric model of *Bacillus subtilis* where its fluxes analyse predicted transcriptional regulation to be more important for the dynamics induced by glucose than by malate [7].

In metabolic pathway analyses, the metabolic networks are described in term of biochemical reactions and metabolites. The integration of structural data is required for a comprehensive understanding of the metabolic networks. We represent the metabolic networks with the functional connectivity between the protein functional domains to make more relevant analyses. We used BioΨ, a formal multi-level description based on elementary actions, to assign functions on structural domains and the elementary flux modes theory to check if the already known pathways remain presents and to identify new ones.

A new version of the software has `Mpas` (Metabolic Pathway Analyser Software) been developed during a Master2'internship by Gh. Fievet. Meanwhile we have also introduced in the landscape of the cell its membranes and the numerous pumps that facilitate ions transfers, hence taking into account the pH of the cytoplasm, a parameter that fits the cell mytosis cycle and which proves to separate the cancerous/normal status of cells [22]. We now aim at study larger and more elaborate metabolic systems, including the Krebs cycle and the mitochondria influence, thus enhancing the scalability of our method [17].

5.4.3. Bacterial phenotypic adaptation

We attempt to re-interpret a major event, the initiation of chromosome replication in *Escherichia coli*, in the light of scales of equilibria. This entails thinking in terms of hyperstructures as responsible for intensity sensing and quantity sensing and how this sensing might help explain the role of the DnaA protein in initiation of replication. We outline experiments and an automaton approach to the cell cycle that should test and refine the scales concept [19].

Another possible direction to study the mechanisms used by cells to integrate and respond to their environment is to search for a link between two large hyperstructures: the cytoskeleton and the general metabolic activity of the cell. There is extensive evidence for the interaction of metabolic enzymes with the eukaryotic cytoskeleton. We state the hypothesis that the cytoskeleton senses and integrates the general metabolic activity of the cell. The physical and chemical effects arising from metabolic sensing by the cytoskeleton would have major consequences on cell shape, dynamics and cell cycle progression. The hypothesis provides a framework that helps the significance of the enzyme-decorated cytoskeleton be determined [18].

In order to test these hypotheses, we have added many features to the HSIM simulation software. The main addition being a way to get both the power of expression of the "entity-centred" paradigm and the computational efficiency of global methods, such that Gillespie-like stochastic simulation algorithm (SSA). To achieve this, we have implemented two new algorithms. The first one concerns the possibility to take into account the interactions between two classes of molecules: the one we want to follow the spatial location over time (entities) and the one for which only the evolution of the number of copies over time is relevant.

The second algorithm is an enhancement of the tau-leap variant of the exact Gillespie SSA; This allows to take into account the interactions between globally treated molecules. The HSIM-SSA algorithm performs an adaptive processing of the number of reactions which may have been triggered during the time step. At each time step, the fast reactions are averaged while the slow reactions are fully stochastically treated. This allows HSIM-SSA to be more than 10 times faster than the other tau-leap SSA implementations [28].

5.4.4. Use of bacteria for biotechnology

Another center of interest has been to find a way to use bacteria as a mean to help us to engineer new biomolecules with specific characteristics. It is sometimes speculated that the equivalent of the polymerase chain reaction might be developed for identification of peptides, proteins or other molecules. Natural amplification systems do exist as in the case of certain autoinducer systems in bacteria. We have been outlined a possible, generic method, *the mimic chain reaction*, for obtaining peptides with 3-D structures that mimic the 3-D structure of their targets. These targets would include a variety of molecules, including proteins. There are therefore two categories of applications: the ability via amplification firstly to detect a known protein or other target at an extremely low concentration and secondly to obtain a set of peptides that mimic the structure of an unknown target and that can be used to obtain a *photofit* [20].

6. Partnerships and Cooperations

6.1. Regional Initiatives

J. M. Steyaertwas the coordinator of RNA-omics Digiteo project, P. Clote (Boston College) being a Digiteo chair until June 2012.

A. Denise is the coordinator of the "Japarin-3D" Digiteo project 2012-2016. This project, in collaboration with PRISM at Versailles, aims to develop new efficient approaches for predicting the 3D structure of large RNA molecules, by applying game theory and graph algorithms.

6.2. National Initiatives

6.2.1. ANR

A. Denise is coordinator of the ANR project AMIS ARN 2009-2012 (ANR-09-BLAN-0160) and is involved in the NSD-NGD ANR project 2010-2014. Y. Ponty is involved in the MAGNUM ANR project (BLAN program, 12/2010–12/2014).

6.3. International Initiatives

6.3.1. Inria Associate Teams

6.3.1.1. ITSNAP

Title: Intelligent Techniques for Structure of Nucleic Acids and Proteins

Inria principal investigator: Julie Bernauer

International Partner (Institution - Laboratory - Researcher):

SLAC National Accelerator Laboratory (United States) - Stanford Synchrotron Radiation Laboratory - Henry van den Bedem

Stanford University (United States) - Computational Structural Biology, School of Medicine, Structural Biology - Michael Levitt

Duration: 2012 - 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. Participation In International Programs

J.Bernauer is coordinator with Pr. X. Huang at the Hong-Kong University of Science and Technology of a Partenariat Hubert Curien (PHC) Procore project (2012-2013). The project is entitled *Computational studies of conformational dynamics of the RNA-induced silencing complex and design of miRNAs to target oncogenes*.

Adrien Rougny, an internship student supervised by C. Froidevaux in AMIB, has successfully applied for the 2nd call of 2012 "NII International Internship Program". In February 2013, he will start an internship at NII for an Internship in Pr. Katsumi Inoue's group on the topic "Inference and Learning for Systems Biology and Network Dynamics".

M. Régnier and D. Sherman (MAGNOME-INRIA) participate to a joint project CARNAGE of France-Russia program.

6.4. International Research Visitors

6.4.1. Visits of International Scientists

R. Giegerich

Institution: Bielefeld University (Germany)

Subject: Efficient algorithms for RNA secondary structure alignment.

Funding: DGAR (Ecole Polytechnique)

R. Giegerich visited the AMIB project-team for a month. He taught dynamic-programming to the students of the BIBS master. He initiated a collaboration on sparsification, an algorithmic technique that speeds up dynamic programming algorithm. A comprehensive review on RNA structure alignment algorithms, to appear in a forthcoming book, was also written during his stay.

J. Waldispühl

Institution: McGill University (Canada)

Subject: RNA design and tertiary structure prediction.

Funding: DIGITEO (LRI)

J. Waldispühl visited AMIB for a month. He finalized a collaboration on RNA design (Y. Ponty, leading to [16]), established a new collaborative research (with A. Denise and Y. Ponty, on tertiary motifs), laid the foundations of a future X-UPSud exchange program, initiated a workshop on molecular interactions (with J. Bernauer), and started a PhD co-supervision (A. Soulé, co-supervised with J.-M. Steyaert and Y. Ponty).

X. Huang

Subject: Millisecond dynamics at atomic resolution by Markov State Models

Institution: Hong Kong University of Science and Technology (Hong-Kong)

A. Sim

Subject: Modeling RNA by hierarchical natural moves

Institution: Stanford University (USA) / A*STAR (Singapore)

L. Pereyaslavets

Subject: Critical assessment of non bonded part of force fields

Institution: Stanford University (USA)

Y. Okamoto

Subject: Protein folding, unfolding, and ligand docking by computer simulations

Institution: Nagoya University (Japan)

6.4.1.1. Internships

A. Martirosyan (March-Jul 2012)

Subject: A Dynamical Model for the Transmembrane Potential Regulation by pH

Institution:Cergy University (Pontoise)

Funding: INRIA

Supervision:L. Paulevé and M. Régnier

B. Brancotte (March-July 2012)

Subject: Designing a framework to compare biological data ranking methods

Institution: Paris-Sud University (France)

Funding: INRIA

Supervision: S. Cohen-Boulakia and A. Denise

Gh.Fievet (March-Sept 2012)
Subject: Improving MPAS software
Institution: Paris-Sud University (France)
Funding: Ecole Polytechnique
Supervision:J.M. Steyaert

J. Weaver (Jun-Aug 2012)
Subject: Efficient Motif Discovery and Evaluation
Institution: Massachusetts Institute of Technology (United States)
Funding: MIT France program
Supervision: Y. Ponty and M. Régnier

A. Menard (Jun-Aug 2012)
Subject: Extending JalView's RNA interconnection with Varna
Institution: Université Paris-Sud
Supervision: Y. Ponty and J. Procter (Univ. Dundee, Scotland)

A. Soulé (Jun-Aug 2012)
Subject: Prediction of RNA-RNA interactions in yeast
Institution: Ecole Polytechnique
Supervision: Y. Ponty and J.-M. Steyaert

V. Arendt (Jun-Aug 2012)
Subject: Integrating RNA web services into JalView using Jabaws
Institution: Duke University (United States)
Funding: Google Summer of Code program
Supervision: Y. Ponty and J. Procter (Univ. Dundee, Scotland)

T. Coulmy & N. Duhamel (Jun-Jul 2012)
Subject: Average-case property analysis of workflows based on hypergraphs
Institution: Université Paris-Sud
Supervision: S. Cohen-Boulakia and Y. Ponty

F.K. Sheong (May-Aug 2012)
Subject: RNA structural design by docking and machine learning
Institution: The Hong Kong University of Science and Technology (Hong Kong, (China))

L. Uroshlev (Oct-Nov 2012)
Subject: Reference state for RNA KB potentials
Institution: IOGEN (Moscou, (Russia))

A. Bari (Oct 2012)
Subject: stress-inducible miRNAs
Institution:El Farabi University (Almaty, (Kazakhstan))

7. Dissemination

7.1. Scientific Animation

7.1.1. French Community

Participants: Patrick Amar, Jérôme Azé, Julie Bernauer, Sarah Cohen-Boulakia, Alain Denise, Christine Froidevaux, Sabine Peres, Yann Ponty, Mireille Régnier, Jean-Marc Steyaert.

The whole team is involved in GDR-BIM (Molecular Bioinformatics, http://www.gdr-bim.u-psud.fr/). J. Azé is the webmaster. Alain Denise is a member of the Scientific Committee. Ch. Froidevaux and S. Cohen-Boulakia participate to the subdomain *Knowledge Representation, Ontologies, Data Integration and Grids*.

A. Denise, Y. Ponty and M. Régnier participate into the subdomain Sequence Analysis and to COMATEGE subgroup of GDR- IM (Informatique Mathématique, http://www.gdr-im.fr/)

A. Denise, Y. Ponty, J.-M. Steyaert, and M. Régnier are involved in the ALEA working group (http://igm.univ-mlv.fr/~nicaud/webalea/ of the GDR-IM (Informatique Mathématique, http://www.gdr-im.fr/).

7.1.2. Seminars and visits

7.1.2.1. Amib seminars

We received in our weekly seminar: A. Thévenin (Bielefeld University), A. Sim (Bioinformatics Institute, A*STAR Singapore), J. Andreani-Feuillet (CEA, Saclay), R. Fonseca (University of Copenhagen), P. Clote (Boston College), J. Waldispuhl (McGill University), Robert Giegerich (Bielefeld Univeristy), J-C. Almeida (Lisbonne), Y. Okamoto (Nagoya University), N. Malod-Dognin (INRIA Sophia-Antipolis), V. Boeva (Institut Curie, Paris), L. Pereyaslavets (Stanford University), X. Huang (HKUST).

7.1.2.2. Other seminars

P. Amar ran a workshop at the *Modelling Complex Biological Systems in the Context of Genomics.* Spring school at Évry. He has been invited to give a talk at the Ecole de Printemps 2012 de la Société Francophone de Biologie Théorique on *Comparative study of some methods for simulation of biochemical reactions*.

J. Azé has been invited to present his work in the field of Protein-Protein Interaction and Docking to the Dalembert's seminars of Paris-Sud.

J. Bernauer gave a talk at the Inria@SiliconValley Workshop BIS2012 in May in Paris. She gave an invited talk at the Computational Structural Biology workshop at HKUST in August (Hong-Kong) at SSRL (Stanford Synchrotron Radiation Lightsource) in November (SLAC, Stanford, USA).

S. Cohen-Boulakia has been invited to participate to the Dagsthul seminar on *Principles on Provenance*. She has been invited to present the current challenges and opportunities in the field of *Data Integration in the Life Sciences* at *Internet Memory*.

A. Denise gave an invited keynote on combinatorics and random generation in the context of bioinformatics at the GASCom 2012 conference (Bordeaux).

S. Peres has been invited to give a talk to the Dagsthul seminar 12462 on *Symbolic Methods for Chemical Reaction Networks*. She gave a talk to *Modelling Complex Biological Systems in the Context of Genomics.* at Évry and to *The 4th* JFLI-LRI-NII *Workshop on Consequence Finding and Satisfiability Testing in Distributed Environments and Systems Biology* at Orsay.

Y. Ponty gave an invited keynote on RNA visualization at the EMBL-hosted conference *VIZBI'12* (Heidelberg, Germany), three invited talks at the *Benasque RNA workshop* (Spain), and an invited talk at the SeqBio'12 colloquium (LIGM, Marne La Vallée, France).

M. Régnier gave an invited talk at the Conference "en l'honneur d' Alain Guenoche", CIRM, Marseille, France.

7.1.2.3. International exchanges

J.Bernauer and M.Régnier visited H. van den Bedem at SSRL (SLAC) and M. Levitt at Stanford University (USA). J.Bernauer visited the Huang group at HKUST (Hong-Kong) and the Bujnicki lab at IIMCB (Warsaw, Poland).

M. Régnier and D. Iakovishina visited IoGene (Moscow). M. Régnier visited USC (M. Waterman) and UC Berkeley (P. Novichkov and E. Purdom).

S. Cohen-Boulakia visited the Information Management Group at the University of Manchester (C. Goble and N. Paton) to initiate collaboration on Data Integration for the Life Sciences (including provenance in scientific workflows).

Katsumi Inoue's group visited AMIB group during the 4th JFLI-LRI-NII Workshop on "Consequence Finding and Satisfiability Testing in Distributed Environments and Systems Biology" (19th - 20th, November 2012).

7.1.3. Program Committee

P. Amar was chairman of the organising committee, and a member of the scientific committee as well, for the conference "Modelling Complex Biological Systems in the context of genomics", Evry, May 2012. (http://epigenomique.free.fr/en).

J. Azé and C. Froidevaux served as PC members for the conference EGC 2012 (Extraction et la Gestion des Connaissances, Bordeaux).

S. Cohen-Boulakia served as PC member for: ICDE 2012 22 (27th I 2012 222tnf. on Data Engineering, Washington, USA), SWEET 2012 (Int. workshop on Scalable Workflow Enactment Engines and Technologies, Scottsdale, USA) and BDA 2012 (Bases de Données Avancées, Clermont-Ferrand).

S. Cohen-Boulakia and C. Froidevaux served as PC member for DILS 2012 (Int. workshop on Data Integration in the Life Sciences, University of Maryland, USA).

A. Denise is a member of the editorial board of Technique et Sciences Informatiques.

F. D'Alche-Buc, C. Froidevaux, and Y. Ponty served as PC members for JOBIM 2012 (Journées Ouvertes en Biologie, Informatique et Mathématiques, Rennes).

Ch. Froidevaux served as member of the program committee of JOBIM'12 bioinformatics conference.
She was PC member of the following international workshops and conferences: CMBS 2012 2012 (25th International Symposium on Computer-Based Medical Systems, Rome, Italy), NETTAB'2012 (12th International Workshop on Network Tools and Applications in Biology, on Integrated Bio-Search, Como, Italy) ICCS 2012 (Workshop on Biomedical and Bioinformatics Challenges to Computer Science, Omaha, USA) and ISMB 2012 (Applied Bioinformatics area, Long Beach, USA).

J. Bernauer, Y. Ponty and M. Régnier served as PC members for WRSBS 2012 (1st International Workshop on Robustness and Stability of Biological Systems and Computational Solutions, Orlando, USA).

Y. Ponty, M. Régnier, and J.-M. Steyaert served as PC members for BICOB 2012 (4th International Conference on Bioinformatics and Computational Biology, Las Vegas, USA).

S. Peres served as PC member for ECCB 2012 (European Conference on Computational Biology 2012, Basel, Switzerland).

7.1.4. Research administration

Y. Ponty is an elected member of the *Comité national du CNRS* (6th section – Foundations of Computer Science and CID 51 –Bioinformatics).

J. Bernauer is member of the IDEX Paris - Saclay Groupe de travail Sciences du Vivant.

A. Denise was a member of the *Comité national du CNRS* (7th section – and CID 43 –Bioinformatics) until September 2012. He is an expert for the *Direction Géenérale de la Recherche et l'Innovation* (DGRI) of the Research Ministry. He is a member of the Scientific Commission of the Inria-Saclay reserach center.

Ch. Froidevaux has been head of the Computer Science Department at the University Paris Sud until the end of January 2012 and is head of the Bioinfo group at LRI.

M. Régnier is a deputy-member of DIGITEO program committee.

J.-M.Steyaert is a member of the Board of Administrators of Polytechnique.

7.2. Teaching - Supervision - Juries

7.2.1. Teaching

The Master of Bioinformatics and Biostatistics, which is a joint master between University Paris-Sud and Ecole Polytechnique http://www.bibs.u-psud.fr, is co-headed by members of the group.

J.-M. Steyaert organizes BIBS (M1 and M2) at Ecole Polytechnique. A. Denise has co-headed the Master (M1 and M2) at the University Paris Sud until end August. C. Froidevaux is co-heading it at the University since September 2012.
Most team members are teaching in this master.

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

Cycle Ingénieur Polytechnicien: J. Bernauer, Modal Bioinformatique, 18h, 2ème année, École Polytechnique, France

Cycle Ingénieur Polytechnicien: J. Bernauer, PSC, encadrement, 2ème année, Ecole Polytechnique, France

Cycle Ingénieur Agro Paris Tech: J. Bernauer, Module AAB, cours invité, 3ème année, Agro Paris Tech, France

Master BIM: Y. Ponty, Modélisation et bioinformatique de l'ARN, 8h de cours, M2, Université Paris-Sud, France

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

Doctorat : M. Régnier, Combinatorics on genome, 20h, El-Farabi University, Kazakhstan

Master : J.-M. Steyaert, X cycle ingénieur INF582- Datamining, 35h, M1, Ecole Polytechnique, France

Licence : J.-M. Steyaert, X cycle ingénieur Modal-BioInformatique, 45h, L3, Ecole Polytechnique, France

Master : J.-M. Steyaert, BIBS Algorithmique avancée et optimisation, 25h, M2, X-Orsay, M2, Ecole Polytechnique, France

Data Bases, 48h, M1 BIBS (Bioinformatics and BioStatistics), Paris-Sud University, France (C. Froidevaux)

Advanced Algorithmics, 48h, M1 BIBS (Bioinformatics and BioStatistics), Paris-Sud University, France (C. Froidevaux)

Integration and Analysis of heterogeneous data from the Web, 24h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (J. Azé, S. Cohen Boulakia, C. Froidevaux)

Advanced Data Bases and Data Mining, 42h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (S. Cohen Boulakia, C. Froidevaux).

Initiation to Research, 6h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University, France (C. Froidevaux)

Software Engineering for Bioinformatics, 48h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (P. Amar)

Modelling and Simulation of Biological Processes, 24h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (P. Amar)

Biological Networks and Systems Biology, 9h, M1 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (P. Amar)

RNAomics and RNA Bioinformatics, 12h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (A.Denise)

Theoretical Computer Science, 30h, M2 BIBS (Bioinformatics and BioStatistics), Paris-Sud University/École Polytechnique, France (A.Denise)

7.2.2. Supervision

HdR : Jérôme Azé, Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs, Paris-Sud University, 16/11/2012

PhD : Feng Lou, Algorithms for studying RNA secondary structures and sequence alignments, Univ. Paris-Sud, 30/01/2012, P. Clote (Boston) and A. Denise.

PhD : Philippe Rinaudo, Algorithmics of RNA structure-sequence alignment: a general and parameterized approach, Univ. Paris-Sud, 05/12/2012, D. Barth (Univ. Versailles) and A. Denise.

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

PhD in progress : Jiuqiang Chen, Mining and Integrating heterogeneous data in e-science environments, 10/09/2011, Université Paris Sud, S. Cohen-Boulakia and C. Froidevaux

PhD in progress: Adrien Guilhot-Gaudeffroy, Modelling and scoring of protein-RNA complexes, 01/10/2011, J. Azé, J. Bernauer, C. Froidevaux

PhD in progress: Daria Iakovishina, A Combinatorial Approach to Assembly Algorithms, 01/11/2011, M. Régnier

PhD in progress : Cécile Pereira, Bioinformatics approaches for a comparative study of metabolic networks and their evolution, 01/10/2011, A. Denise and O. Lespinet (IGM, Univ. Paris-Sud)

PhD in progress : Antoine Soulé, Evolutionary study of RNA-RNA interactions in yeast, 01/09/2012, J.-M. Steyaert, Y. Ponty, and J. Waldispühl (University McGill, Canada)

PhD in progress : Bo Yang, Bioinformatics approaches for studying the relations between RNA structure and pre-messenger RNA splicing, 01/10/2011, A. Denise and Fu Xiangdong (Wuhan University, China)

PhD in progress : Cong Zeng, Identification of structural motifs in messenger RNAs, 01/10/2011, A. Denise

PhD in progress : Mélanie Boudard, Game theory and stochastic learning for predicting the three-dimensional structure of large RNA molecules , 15/10/2012, J. Cohen (CNRS, Univ. Versailles) and A. Denise.

7.2.3. Juries

A. Denise served as referee and jury member for Sylvain Sené's HDR defence (Univ. Evry, November 2012), Amine Ghozlane's PhD defence (Univ. Bordeaux I, December 2012) and Romain Pogorelcnik's PhD Defence (Univ. Blaise Pascal - Clermont II, December 2012). He served as the head of the jury member in Leandro Montero's PhD defence (Univ. Paris-Sud, December 2012). He served as a jury member in the hiring committee for a *Maître de Conférence sur chaire CNRS* position at Ecole Centrale de Nantes.

C. Froidevaux was a referee for Johan Estellon's PhD defence(Univ. Grenoble). She served as a member of the *Comité de thèse* of Pauline Gloaguen (INRA Tours, Feb 2012) and as a jury member for the defence of her PhD (Univ. Tours, Dec 2012). She served as the head of the jury in Konstantinos Karanasos' PhD defence (Paris Sud, July 2012). She was member of the *HDR* de Jérôme Azé (Univ. Paris-Sud).

Ch. Froidevaux took part as an external jury member in the hiring committee of an Assistant Professor position at IGM (Univ. Paris Sud, Orsay). She is member of the *Commission Consultative de Spécialistes* (CCSU, 27th section) of the University Paris Sud.

M. Régnier served as a jury member in Jérémie Bourdon's HDR defence (Nantes University). She was a member of the hiring committee for *Maître de Conférence* position at Bordeaux University (LaBRI).

S. Peres served as a jury member in the PhD defense committee of Chamseddine KIFAGI at Sfax University (Tunisia).

Y. Ponty served as a jury member in the hiring committee for a *Maître de Conférence* position at Université de Versailles St Quentin. He served as a jury member in the PhD defense committee of Azadeh Saffarian at Université de Lille 1 (LIFL/Inria Lille).

7.3. Popularization

AMIB animated the INRIA booth, at the yearly "Nuit des Chercheurs" http://events.polytechnique.fr/accueil/la-nuit-des-chercheurs/la-nuit-des-chercheurs-43774.kjsp in September at Ecole Polytechnique. The topic was to illustrate the principles underlying RNA folding algorithms through playing combinatorial games.

Y. Ponty gave a presentation at *Unithé ou Café*, the monthly *popular science* event of sc Inria Saclay, on RNA folding using dynamic programming. He was interviewed by the *Interstices* web site on RNA folding algorithms, leading to a 12 minutes-long podcast. Y. Ponty was invited to give a tutorial on RNA visualization at the EMBL (Heidelberg, Germany)-hosted conference VIZBI'12.

S. Cohen-Boulakia was invited to give a short tutorial on *Data Integration in the Life Sciences* at the national thematic school *Masses de données distribuées* (Summer School).

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] 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], http://hal.inria.fr/hal-00483581/en

[3] 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, in press [*DOI :* 10.1016/J.JMB.2011.09.031], http://hal.inria.fr/inria-00637848

[4] 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

Doctoral Dissertations and Habilitation Theses

[5] J. AZÉ. , *Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs*, Université Paris Sud - Paris XI, November 2012, HDR, http://hal.inria.fr/tel-00763947

Articles in International Peer-Reviewed Journals

[6] J. ALLALI, C. SAULE, C. CHAUVE, Y. D'AUBENTON-CARAFA, A. DENISE, C. DREVET, P. FERRARO, D. GAUTHERET, C. HERRBACH, F. LECLERC, A. DE MONTE, A. OUANGRAOUA, M.-F. SAGOT, M. TERMIER, C. THERMES, H. TOUZET. *BRASERO: A resource for benchmarking RNA secondary structure comparison algorithms*, in "Advances in Bioinformatics", 2012, 5 p. , Epub 2012 May 23 [*DOI :* 10.1155/2012/893048], http://hal.inria.fr/hal-00647725

[7] J. M. BUESCHER, W. LIEBERMEISTER, M. JULES, M. UHR, J. MUNTEL, E. BOTELLA, B. HESSLING, R. J. KLEIJN, L. LE CHAT, F. LECOINTE, U. MÄDER, P. NICOLAS, S. PIERSMA, F. RÜGHEIMER, D. BECHER, P. BESSIERES, E. BIDNENKO, E. L. DENHAM, E. DERVYN, K. M. DEVINE, G. DOHERTY, S. DRULHE, L. FELICORI, M. J. FOGG, A. GOELZER, A. HANSEN, C. R. HARWOOD, M. HECKER, S. HUBNER, C. HULTSCHIG, H. JARMER, E. KLIPP, A. LEDUC, P. LEWIS, F. MOLINA, P. NOIROT, S. PERES, N. PIGEONNEAU, S. POHL, S. RASMUSSEN, B. RINN, M. SCHAFFER, J. SCHNIDDER, B. SCHWIKOWSKI, J. M. VAN DIJL, P. VEIGA, S. WALSH, A. J. WILKINSON, J. STELLING, S. AYMERICH, U. SAUER. *Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism.*, in "Science", March 2012, vol. 335, n^{o} 6072, pp. 1099-103 [*DOI :* 10.1126/SCIENCE.1206871], http://hal.inria.fr/hal-00756388

[8] P. CLOTE, Y. PONTY, J.-M. STEYAERT. *Expected distance between terminal nucleotides of RNA secondary structures.*, in "Journal of Mathematical Biology", September 2012, vol. 65, n^{o} 3, pp. 581-99 [*DOI :* 10.1007/s00285-011-0467-8], http://hal.inria.fr/inria-00619921

[9] D. DE VIENNE, J. AZÉ. *Efficient Prediction of Co-Complexed Proteins Based on Coevolution*, in "PLoS ONE", November 2012, http://hal.inria.fr/hal-00756120

[10] 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

[11] A. DENISE, P. RINAUDO. *Optimisation problems for pairwise RNA sequence and structure comparison: a brief survey*, in "Transactions on Computational Collective Intelligence", 2012, to appear, http://hal.inria.fr/hal-00759573

[12] 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^{o} 4, pp. 395-405 [*DOI :* 10.1093/BIB/BBR077], http://hal.inria.fr/hal-00684530

[13] A. ISSABEKOVA, O. BERILLO, M. REGNIER, A. IVASHCHENKO. *Interactions of intergenic microRNAs with mRNAs of genes involved in carcinogenesis*, in "Bioinformation", June 2012, vol. 8, n^{o} 11, pp. 513-518 [*DOI :* 10.6026/97320630008513], http://hal.inria.fr/hal-00736694

[14] 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

[15] A. LAMIABLE, F. QUESSETTE, S. VIAL, D. BARTH, A. DENISE. *An algorithmic game-theory approach for coarse-grain prediction of RNA 3D structure*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2012, page : to appear, http://hal.inria.fr/hal-00756340

[16] 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

[17] D. MORVAN, J. M. STEYAERT, L. SCHWARTZ, M. ISRAËL, A. DEMIDEM. *Normal human melanocytes exposed to chronic insulin and glucose supplementation undergo oncogenic changes and methyl group metabolism cellular redistribution.*, in "AJP - Endocrinology and Metabolism", June 2012, vol. 302, n° 11, pp. E1407-18 [*DOI :* 10.1152/AJPENDO.00594.2011], http://hal.inria.fr/hal-00761765

[18] V. NORRIS, P. AMAR, G. LEGENT, C. RIPOLL, M. THELLIER, J. OVADI. *Sensor potency of the moonlighting enzyme-decorated cytoskeleton*, in "BMC Biochemistry", 2012, to appear, http://hal.inria.fr/hal-00766058

[19] V. NORRIS, P. AMAR. *Chromosome Replication in Escherichia coli: Life on the Scales*, in "Life", October 2012, vol. 2, n° 4, pp. 286-312 [*DOI :* 10.3390/LIFE2040286], http://hal.inria.fr/hal-00758130

[20] V. NORRIS, A. R. THIERRY, P. AMAR, B. I. HOLLAND, F. MOLINA. *The Mimic Chain Reaction*, in "Journal of Molecular Microbiology and Biotechnology", December 2012, vol. 22, n° 5, pp. 335-343 [*DOI :* 10.1159/000345328], http://hal.inria.fr/hal-00761608

[21] 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", July 2012, page : to appear, http://hal.inria.fr/hal-00716558

[22] L. SCHWARTZ, A. GUAIS, M. ISRAËL, B. JUNOD, J.-M. STEYAERT, E. CRESPI, G. BARONZIO, M. ABOLHASSANI. *Tumor regression with a combination of drugs interfering with the tumor metabolism: efficacy of hydroxycitrate, lipoic acid and capsaicin.*, in "Investigational New Drugs", July 2012, epub ahead of print [*DOI :* 10.1007/S10637-012-9849-Z], http://hal.inria.fr/hal-00761761

[23] S. SHEIKH, P. CHASSIGNET, J.-M. STEYAERT, T. V. D. TRAN. *A graph-theoretic approach for classification and structure prediction of transmembrane beta-barrel proteins*, in "BMC Genomics", 2012, http://hal.inria.fr/hal-00650429

[24] 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° 2, 1241010 p. [*DOI :* 10.1142/S0219720012410107], http://hal.inria.fr/hal-00757761

[25] V. D. T. TRAN, P. CHASSIGNET, J.-M. STEYAERT. *Supersecondary structure prediction of transmembrane beta-barrel proteins.*, in "Methods in Molecular Biology -Clifton then Totowa-", 2013, vol. 932, pp. 277-94 [*DOI :* 10.1007/978-1-62703-065-6_17], http://hal.inria.fr/hal-00761759

Invited Conferences

[26] Y. PONTY. *Visualizing 2D & 3D Structures of RNA*, in "VIZBI - 3rd international meeting on Visualizing Biological Data - 2012", Heidelberg, Germany, Jim, Procter (U-Dundee, Scotland) and Sean O'Donoghue (CMIS, Australia), March 2012, http://hal.inria.fr/hal-00723351

[27] J. WALDISPÜHL, B. BERGER, S. DEVADAS, P. CLOTE, Y. PONTY. *Efficient Algorithms to Explore the RNA Mutational Landscape*, in "DM - SIAM Conference on Discrete Mathematics - 2012", Halifax, Canada, SIAM Activity Group on Discrete Mathematics, 2012, http://hal.inria.fr/hal-00723333

International Conferences with Proceedings

[28] 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

[29] 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, United Kingdom, September 2012, 103-110, http://hal.inria.fr/hal-00748041

[30] C. BANDERIER, O. BODINI, Y. PONTY, H. TAFAT. *On the diversity of pattern distributions in combinatorial systems*, in "ANALCO'12", Kyoto, Japan, Omnipress, 2012, pp. 107–116, http://hal.inria.fr/hal-00643598

[31] J. CHEN, C. FROIDEVAUX, C. GOBLE, A. WILLIAMS, S. COHEN-BOULAKIA. *Distilling Structure in Scientific Workflows*, in "Proc. of the 12th International Workshop on Network Tools and Applications in Biology, Nettab 2012 (Poster)", Como, Italy, Supp B, EMBnet.journal, November 2012, vol. 18, pp. 10-102, http://hal.inria.fr/hal-00748035

[32] S. COHEN-BOULAKIA, C. FROIDEVAUX, J. CHEN. *Scientific Workflow Rewriting while Preserving Provenance*, in "Proc. of the 8th IEEE International Conference in eScience", Chicago, United States, IEEE, October 2012, http://hal.inria.fr/hal-00748032

[33] J. DU BOISBERRANGER, D. GARDY, Y. PONTY. *The weighted words collector*, in "AOFA - 23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms - 2012", Montreal, Canada, F. NICOLAS, C. LUC (editors), DMTCS Proceedings, DMTCS, 2012, vol. dmAQ01, pp. 243–264, http://hal.inria.fr/hal-00666399

[34] S. JANSSEN, L. PAULEVÉ, Y. PONTY, B. RAMAN, M. ZYTNICKI. *Can Probabilistic Model Checking Explore Ribo-Nucleic Acid Folding Space?*, in "IWBDA - 4th International Workshop on Bio-Design Automation - 2012", San Francisco, United States, 2012, http://hal.inria.fr/hal-00712557

[35] P. RINAUDO, Y. PONTY, D. BARTH, A. DENISE. *Tree decomposition and parameterized algorithms for RNA structure-sequence alignment including tertiary interactions and pseudoknots*, in "WABI - 12th Workshop on Algorithms in Bioinformatics - 2012", Ljubljana, Slovenia, B. RAPHAEL, J. TANG (editors), University of Ljubljana, 2012, http://hal.inria.fr/hal-00708580

[36] B. ROMAIN, V. LETORT, O. LUCIDARME, F. D'ALCHÉ-BUC, L. ROUET. *Optimisation of reconstruction for the registration of CT liver perfusion sequences*, in "SPIE Medical Imaging 2012", San Diego, CA, United States, Progress in Biomedical Optics and Imaging, February 2012, vol. 8314, (elec. proc) [*DOI :* 10.1117/12.910929], http://hal.inria.fr/hal-00744964

[37] S. SHEIKH, R. BACKOFEN, Y. PONTY. *Impact Of The Energy Model On The Complexity Of RNA Folding With Pseudoknots*, in "CPM - 23rd Annual Symposium on Combinatorial Pattern Matching - 2012", Helsinki, Finland, J. KÄRKKÄINEN, J. STOYE (editors), Springer, 2012, vol. 7354, pp. 321–333 [*DOI :* 10.1007/978-3-642-31265-6_26], http://hal.inria.fr/hal-00670232

[38] J. STARLINGER, S. COHEN-BOULAKIA, U. LESER. *(Re)Use in Public Scientific Workflow Repositories*, in "Scientific and Statistical Database Management - 24th International Conference, SSDBM 2012", Chania,

Greece, A. AILAMAKI (editor), LNCS, Springer, 2012, vol. 7338, pp. 361-378, http://hal.inria.fr/hal-00748029

National Conferences with Proceedings

[39] S. COHEN-BOULAKIA, C. FROIDEVAUX, J. CHEN. *Réécriture de workflows scientifiques et provenance*, in "Proc. of the 28th Journées de Bases de Données Avancées", Clermont Ferrand, France, 2012, http://hal.inria.fr/hal-00748031

[40] F. JOSSINET, A. LAMIABLE, P. RINAUDO, L. AL-SHIKHLEY, F. QUESSETTE, S. VIAL, D. BARTH, E. WESTHOF, A. DENISE. *Graph Algorithms and Software Framework for Interactive RNA Structure Modelling*, in "Journees Ouvertes de Biologie Informatique et Mathematiques - JOBIM'12", Rennes, France, 2012, pp. 475-476, http://hal.inria.fr/hal-00756345

Conferences without Proceedings

[41] J. BOURDON, M. REGNIER. *Large deviation properties for patterns*, in "LSD&LAW 2012", Londres, United Kingdom, T. LECROQ (editor), Simon J. Puglisi and Golnaz Badkobeh, November 2012, Presentation at LSD&LAW 2012. Submitted., http://hal.inria.fr/hal-00758251

Research Reports

[42] A. MARTIROSYAN, L. PAULEVÉ, C. POIGNARD, M. RÉGNIER, J.-M. STEYAERT, L. SCHWARTZ. , *A Dynamical Model for the Transmembrane Potential Regulation by pH*, Inria, September 2012, http://hal.inria.fr/hal-00769446

Other Publications

[43] Y. PONTY. , *Rule-weighted and terminal-weighted context-free grammars have identical expressivity*, unpublished, http://hal.inria.fr/hal-00693600

References in notes

[44] 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 [*DOI :* 10.1093/BIOINFORMATICS/BTR670]

[45] 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

[46] M. F. LENSINK, S. J. WODAK. *Docking and scoring protein interactions: CAPRI 2009.*, in "Proteins", 2010, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=20806235

[47] 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

[48] 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

[49] D. W. RITCHIE. *Recent progress and future directions in protein-protein docking.*, in "Curr Protein Pept Sci", 2008, vol. 9, n^{o} 1, pp. 1-15, http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?cmd=prlinks&dbfrom=pubmed&retmode=ref&id=18336319