IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Ecole Polytechnique

Activity Report 2013

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, from Oct 2013]
Yann Ponty [CNRS, Researcher]

Faculty Members

Patrick Amar [Univ. Paris-Sud, Associate Professor, HdR]
Alain Denise [Univ. Paris-Sud, Professor, HdR]
Jérôme Azé [Univ. Paris-Sud, Associate Professor, until Aug 2013]
Sarah Cohen-Boulakia [Univ. Paris-Sud, Associate Professor]
Christine Froidevaux [Univ. Paris-Sud,Professor, HdR]
Sabine Pérès [Univ. Paris-Sud, Associate Professor]
Jean-Marc Steyaert [Ecole Polytechnique, HdR]

External Collaborator

Philippe Chassignet [Ecole Polytechnique, Associate Professor]

PhD Students

Erwan Bigan [Ecole Polytechnique]
Mélanie Boudard [Univ. Versailles and Univ. Paris-Sud]
Bryan Brancotte [Univ. Paris-Sud]
Jiuqiang Chen [Univ. Paris-Sud]
Adrien Guilhot-Gaudeffroy [Univ. Paris-Sud]
Daria Iakovishina [Ecole Polytechnique]
Adrien Rougny [ENS Lyon, from Sep 2013]
Antoine Soulé [Ecole Polytechnique, from Oct 2013]
Cong Zeng [Univ. Paris-Sud]
Bo Yang [Univ. Paris-Sud and Wuhan University]

Post-Doctoral Fellow

Rasmus Fonseca [Inria]

Visiting Scientist

Vladimir Reinharz [PhD student, from Jan 2013 until May 2013]

Administrative Assistant

Évelyne Rayssac [Ecole Polytechnique]

2. Overall Objectives

2.1. Introduction

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.

2.2. Highlights of the Year

Michael Levitt, our international collaborator of the ITSNAP Associated team, was awarded the Nobel Prize in Chemistry *for the development of multiscale models for complex chemical systems*. The Nobel lecture is available at http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/levitt-lecture.html.
The *Best application paper* at EGC 2013 was awarded to [34].

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 [29], [23], and the efficient detection of multi-stable RNAs (riboswitches) [74],[32].

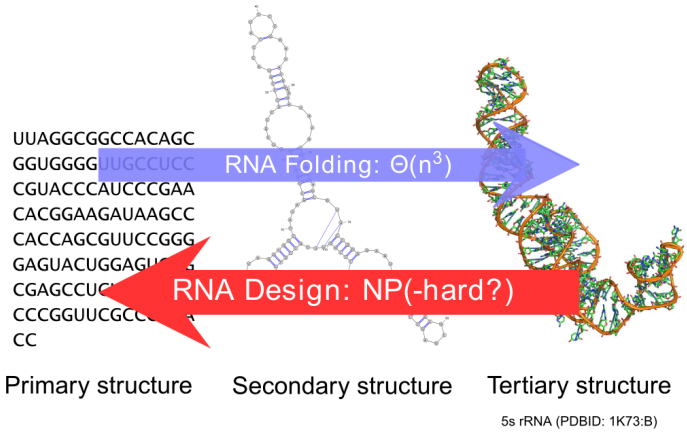

Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

3.1.2. RNA design.

Participants: Alain Denise, 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 [57] 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 [64].

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 [78], [54], [68], [67]. 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 [66]. For RNA, progress in this direction has been realized as well. A kinematics-based conformational sampling algorithm, KGS, for loops was recently developed [62], 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 [62], [78] 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], [75] 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: Julie Bernauer, Alain Denise, Mireille Régnier, Yann Ponty, Jean-Marc Steyaert, 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 [69]. 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, 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 [44]. 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 [70] 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 J. Moussu).

As a second 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. Third, Next Generation Sequencing open the way to the study of structural variants in the genome, as recently described in [51]. 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 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 [58]: 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 [71], [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 [69].

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 [56],[19].

3.3. Geometry and machine learning for 3D interaction prediction

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

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 [50], [60] and efficient code are essential [65].

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 [59]. 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 [48] 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 [76], 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 [76].

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 [59]. 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 [49]. 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 [52], 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 [53].

We also decided to extend these techniques to the analysis of protein-nucleic acid complexes. The first preliminary developments and tests are performed by A. Guilhot (See figure 2).

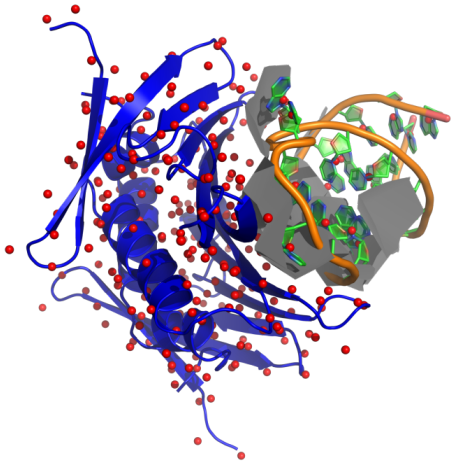

Figure 2. Coarse-grained representation and Voronoi interface model of a PP7 coat protein bound to an RNA hairpin (PDB code 2qux). The Voronoi model captures the features of the interactions such as stacking, even at the coarse-grained level.

3.4. 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: Christine Froidevaux, Sarah Cohen-Boulakia, 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*.

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, Laurent Schwartz, 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. Our approach to this problem classifies into a few classes elementary flux modes which share a set of common reactions, called common motifs.

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 [47].

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, Laurent Schwartz, 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 [45]. 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 [46].

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. Software and Platforms

4.1. VARNA

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) [55]...
VARNA currently ships in its 3.9 version, and consists in ∼50 000 lines of code in ∼250 classes.
Impact: Downloaded ∼10 000 times and is cited by more than ∼170 research manuscripts (source: Google Scholar).
Availability: Distributed under the terms of the GPL v3 licence since 2009 on simple demand to the author(s) at http://varna.lri.fr.

4.2. 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 [63]. More than 300 visits since its release in January 2012.

4.3. Rna3Dmotif

Participant: Alain Denise [correspondant].

Rna3Dmotif is a free bundle of three easy-to-install programs aimed to be used in combination to automatically extract recurrent RNA local tertiary motifs. The approach used is based on a graph representation of the RNA tertiary structure using LW nomenclature. It was applied to several widely studied ribosomal RNA structures and the motifs thus found were deposited in a dedicated repository.
Impact: Cited in 17 research manuscripts (source: Google Scholar).
Availability: Distributed under the terms of the licence since 24/03/2009 on simple demand to the author(s) at http://rna3dmotif.lri.fr.

4.4. GenRGenS

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... [72] GENRGENS currently ships in its 2.0 version, and consists in ∼25 000 lines of code in ∼120 Java classes.
Impact: Downloaded ∼5 000 times and is cited by more than ∼50 research manuscripts (source: Google Scholar).

Availability: Distributed under the terms of the GPL v3 licence since 2006 on simple demand to the author(s) at https://www.lri.fr/ genrgens/.

4.5. 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.

4.6. 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.

4.7. 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. SPFlow

Participant: Sarah Cohen-Boulakia [correspondant].

Scientific workflow systems are numerous and equipped of provenance modules able to collect data produced and consumed during workflow runs to enhance reproducibility. An increasing number of approaches have been developed to help managing provenance information. Some of them are able to process data in a polynomial time but they require workflows to have series-parallel (SP) structures. Rewriting any workflow into an SP workflow is thus particularly important.
SPFLOW answers this need and takes in a workflow (from the Taverna system) and provide a runnable and provenance equivalent (Taverna) workflow."
Impact: The tool is currently used by Taverna's users from the University of Manchester and more generally by myExperiment users.
Availability: Distributed under the terms of the licence since 04/02/2013 on simple demand to the author(s) at http://www.lri.fr/ chenj/SPFlow/.

4.9. SPChecker

Participant: Sarah Cohen-Boulakia [correspondant].

Scientific workflow systems are numerous and equipped of provenance modules able to collect data produced and consumed during workflow runs to enhance reproducibility. An increasing number of approaches have been developed to help managing provenance information. Some of them are able to process data in a polynomial time but they require workflows to have series-parallel (SP) structures.
SPChecker is able to detect whether or not any Taverna workflow has a series-parallel structure.
Impact: The tool is currently used by Taverna's users from the University of Manchester and more generally by myExperiment users (a collaboration with Manchester has started and should significantly augment the number of potential users).
Availability: Distributed under the terms of the licence since 01/02/2013 on simple demand to the author(s) at http://www.lri.fr/ chenj/SPChecker/.

4.10. BioGuide

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

BioGuide/BioGuideSRS : this software helps the scientists choose suitable sources and tools, find complementary information in sources, and deal with divergent data.
Reference : Sarah Cohen-Boulakia, Olivier Biton, Susan Davidson, Christine Froidevaux, BioGuideSRS: Querying Multiple Sources with a user-centric perspective, Bioinformatics, March, 23(10), 1301-1303, 2007.
Impact: The paper related to the tool has been cited by $\sim$26 research manuscripts (source: Google Scholar) so far. Since 2007 and up to now, BioGuide has 8,030 distinct users including regular users from the EBI (European Bioinformatics Institute), the Institut Curie and the Children's Hospital of Philadelphia.
Availability: Distributed under the terms of the licence since 01/09/2006 on simple demand to the author(s) at http://bioguide-project.net/.

4.11. 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.

5. New Results

5.1. RNA

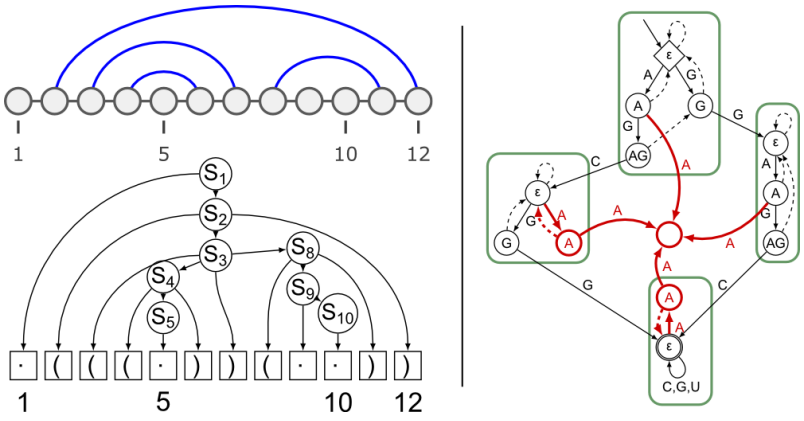

Figure 3. Language-theoretical constructs for the constrained design of RNAs. Starting from a secondary structure, the language of sequences compatible with base-pairing constraints is modeled as a context-free grammar (Left), while forced and forbidden motifs (here, $\{\mathsf{AA}\}$ is forbidden, and $\{\mathsf{AGC}, \mathsf{GG}\}$ are forced) can be modeled by a dedicated automaton (Right).

5.1.1. RNA design through random generation

Extensive experiments revealed a drift of existing software towards sequences with a high G+C-content. Relying on our random generation methods, we showed how to control this distributional bias in sequences using a multidimensional Boltzmann sampling [30], [22]. We also explored the combination of random generation (global sampling) and local search into a novel category of *glocal* approaches, yielding promising results.

Finally, we explored language-theoretic constructs, namely products of finite-state automata and context-free languages, to force or forbid the presence of identified functional motifs within designed sequences [33].

5.1.2. Towards 3D modeling of large molecules

Ab initio research benefited from our works on research and classification of RNA structural motifs [63]. Significant progress towards the *ab initio* prediction of the 3D structure of large RNAs were achieved. This problem is beyond the scope of current approaches and we proposed a promising coarse-grained approach based on game theory [13] that scales up to several hundreds of bases.

5.1.3. Fast-fourier transform for riboswitch

In the field of RNA computational biology, many algorithms use dynamic programming to partition the folding landscape according to a set of structural parameters. More precisely, the goal is to compute the number (resp. cumulated Boltzmann weight) $c_{p_1,p_2,p_3\ldots}$ of secondary structures having p_i occurrences of some structural parameter P_i, where P_i may denote the distance to a reference structure, the number of # helices, base-pairs...The resulting algorithms, although polynomial in theory, are usually unusable in practice, particularly due to their unreasonable complexities (typically $\Theta(n^{3+2k})/\Theta(n^{2+k})$ time/memory for k parameters) and the intrinsic difficulties one encounters while trying to distribute their computation over multiple processors (highly connected dependency graph).

In collaboration with P. Clote's group (Boston College), we have described generic algorithmic principles to dramatically decrease these complexities, and make this class of algorithms practical. The main idea is to capture the partitioned space within a large polynomial, which can typically be efficiently evaluated (typically in $\Theta(n^3)$) as soon as the parameters are additive. One can then perform (possibly in parallel) $\Theta(n^k)$

independent evaluations of the polynomial, and use the Discrete Fourier Transform to recover the coefficients in $\Theta(k \cdot n^k \cdot \log(n))$ time. Applying these principles to the `RNAbor` algorithm, whose complexities were in $\Theta(n^5)/\Theta(n^3)$, we obtained an novel $\Theta(n^4)/\Theta(n^2)$ (parallelizable in $\Theta(n^3)/\Theta(n^2)$ time/memory on $m \to \infty$ processors), we obtained a novel algorithm to detect bistable thermodynamic structures, such as riboswitches, which we presented at `Recomb`'13 [32].

5.2. Sequences

5.2.1. Random generation

The random generation of decomposable combinatorial structures, pioneered by P. Flajolet in the 80s, provides an elegant, yet powerful, framework to model and sample the objects which appear in computational biology. Random samples can then be used to assert the significance of a given observable when closed form formulae are difficult to obtain.

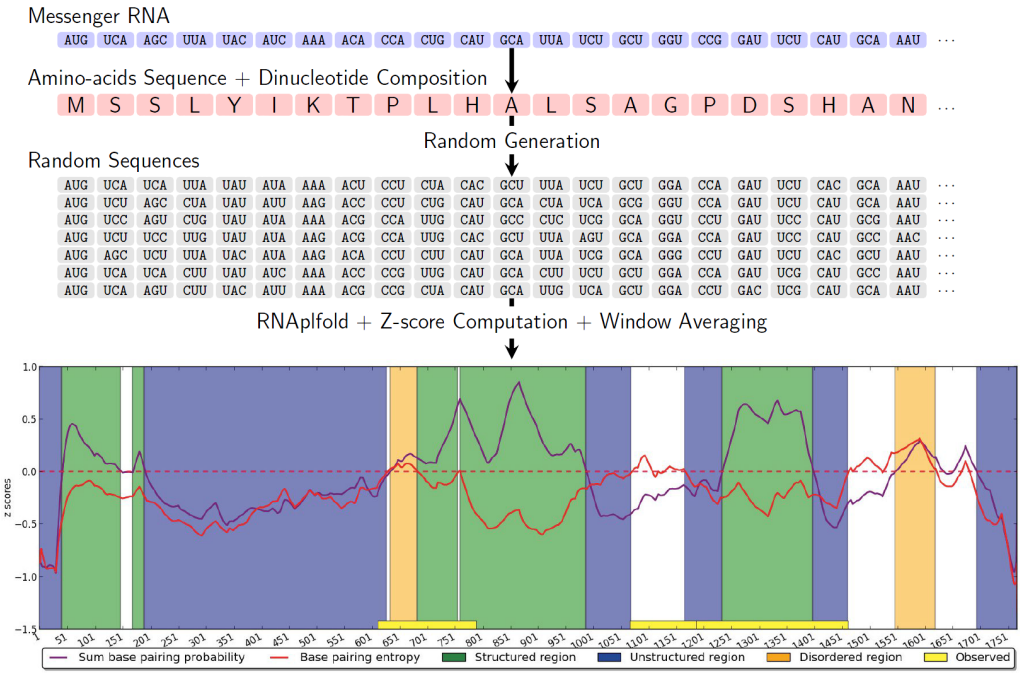

Figure 4. Workflow of our `NASP` pipeline [27]: An assessment of significantly (un)-structured regions in protein-coding RNAs can be achieved through a dinucleotide-preserving random generation of sequences encoding the same protein.

Messenger RNAs (mRNAs) encode proteins, but may also independently feature structured motifs which are crucial to recoding and alternative splicing mechanisms. In order to predict such motifs, the stability of smaller regions within a given mRNA must be compared to that of sequences generated with respect to a **background model** which, at the same time, preserves the encoded amino-acid sequence and the capacity of the overall sequence to form a stable fold (proxy-ed by the dinucleotide composition). Using multidimensional Boltzmann sampling, we have revisited the underlying – well-defined, yet never solved exactly – random generation problem, and provided the first unbiased and practical algorithm for the problem [27]. The algorithm, developed in collaboration with McGill and Université de Montréal (Canada), has linear time complexity as soon as a small tolerance (typically $\Theta(1/\sqrt{n})$) on the composition is allowed.

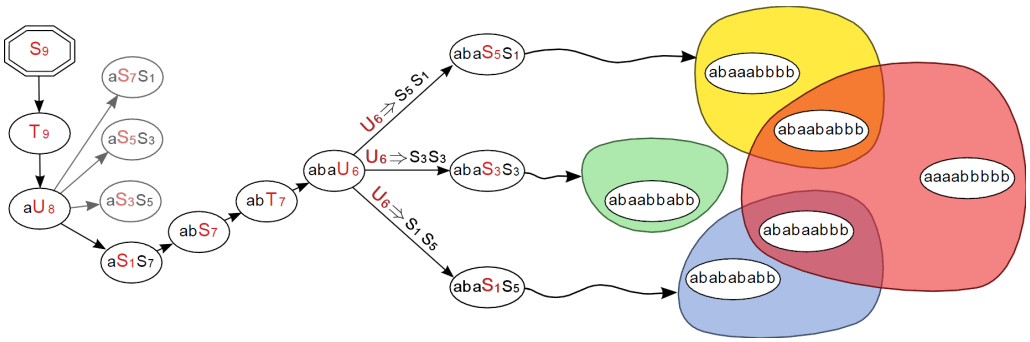

Figure 5. A uniform random generation of words avoiding a predefined set of words can be achieved using a dedicated data structure, leading to a careful correction of the emission probabilities. Enriching the set of forbidden words after each generation, one obtains a non-redundant generation algorithm [17].

Some other biological objects, such as RNA secondary structures, naturally appear with probabilities which are poorly modeled by the uniform distribution. To better model such objects, Denise *et al* [3] have introduced the **weighted distribution**, and adapted classic random generation algorithms such that each object within a given combinatorial family can be generated with respect to it. However, the exponentially increasing probability ratio between the most and least probable object sometimes leads to a large degree of redundancy within generated sets . To work around this issue, and generate non-redundant sets of objects, we have proposed a sequential algorithm with deterministically avoids any previously generated word, without introducing any bias in the generation [17].

Besides, in collaboration with the Fortesse group at LRI, we developed a new divide and conquer algorithm for the random generation of words of regular languages, and we performed a complete benchmarking of all state-of-the-art methods dedicates to this problem [56].

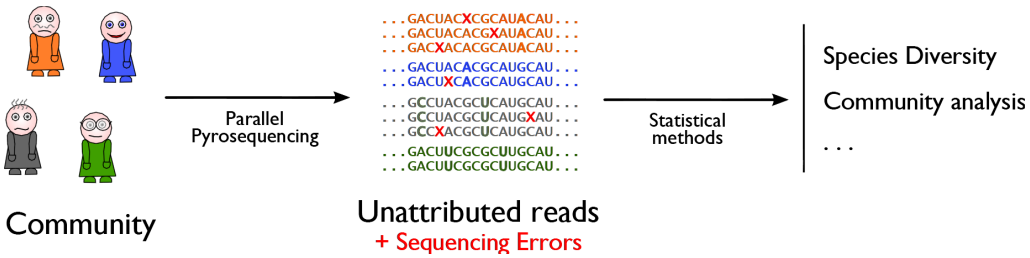

Figure 6. While simultaneously sequencing the genome of a (microbial) community, Next-Generation Sequencing techniques produce small genomic fragments, whose diversity arises from a combination of genetic variants and sequencing errors. We used knowledge of the RNA secondary structure to develop a pre-filter that detects and corrects post-mapping sequencing errors.

5.2.2. Next Generation Sequencing (NGS)

As a side-product of our previous collaborative studies with J. Waldispühl (McGill, Canada), focusing on sequence/structure relationship in RNA, we revisited the problem of detecting and correcting RNA sequences

obtained using pyrrosequencing techniques. Indeed, ribosomal RNAs are often used to estimate the population diversity within a microbiome, and sequencing errors may lead to biased estimates. In this context, we investigated whether a complete knowledge of the RNA secondary structure could be exploited to detect and correct errors in NGS reads.

To that end, we introduced a probabilistic model, defined over all sequences at maximal distance d from the input read and their respective folding. This model captures both the stability of the induced fold and its compatibility with a reference multiple sequence alignment. We designed a linear-time inside/outside algorithm to compute exactly the probability that a given position is mutated in the ensemble. Our initial implementation, presented at RECOMB'13 [29] and published an extended version in *Journal of Computational Biology* [23], revealed encouraging results, and we plan to combine it with a population diversity estimator to test its potential in a metagenomics context.

5.2.3. Combinatorics of motifs

An algorithm for pvalue computation has been proposed in [44] 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 [28] 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.

Large deviations approximations are needed for very rare events, e.g. very small pvalues, as Gaussian approximations are known not to be applicable. In [21], combinatorial properties of words allow to provide an explicit and tractable formula for the tail distribution with a low space and time complexity and a guaranteed tightness. Double strands counting problem is addressed where dependencies between a sequence and its complement plays a fundamental role. A large deviation result is also provided for a set of small sequences, with non-identical distributions. 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 a recent internship at LIX, F. Pirot detected a Chi-like motif in Archae genome.

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 [8], [9].

5.3. 3D Modelling and Interactions

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. Therefore, 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. Although this representation may disregard some of the fine details of the molecule conformation, it still retains the general architecture of molecules, and is especially useful in the study of RiboNucleic Acids (RNAs) and transmembrane beta-barrel proteins (TMB). The latter class of proteins accounts for 20 to 30% of identified proteins in a genome but, due to difficulties with standard experimental techniques, they constitute only 2As 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 TMBs are strongly structured objects, model based methodologies [26], [25] are an interesting alternative to these conventional methods. The efficiently obtained 3D structures provide a good model for further 3D and interaction analyses.

In a recent work [34], we focused on the identification of protein-protein complexes based on the putative interaction between pairs of proteins as the sole source of information. From the results obtained on *E. coli*, we started working on the prediction of multi-body protein complexes from sequence information alone.

In our protein-RNA project, we managed to obtain the first learning results. We optimized the RosettaDock scores and showed that such an optimization cannot be done efficiently without expert knowledge. The first results are to be presented at EGC in 2014 [61].

5.3.1. Large scale cross-docking study of the specificity of protein-protein interactions

The year 2013 saw the conclusion of a long-term collaboration, involving A. Carbone (UPMC) and A. Lopes (IGM, Paris XI). In a recent paper published in the prestigious *Plos Computational Biology* [16] journal, we showed that combining coarse-grain molecular cross-docking simulations and binding site predictions based on evolutionary sequence analysis is a viable route to identify true interacting partners for hundreds of proteins with a variate set of protein structures and interfaces. Also, we realized a large-scale analysis of protein binding promiscuity and provided a numerical characterization of partner competition and level of interaction strength for about 28000 false-partner interactions. Finally, we demonstrated that binding site prediction is useful to discriminate native partners, but also to scale up the approach to thousands of protein interactions. This study was based on a large computational effort made by thousands of internet users helping the World Community Grid over a period of 7 months.

5.4. Data Integration

Work performed in the Data Integration axis this year has been dedicated to the design and implementation of a new approach to reduce the complexity of scientific workflow structures. More precisely, we focused on the presence of "anti-patterns" in the workflow structures, idiomatic structures that lead to over-complicated design. We have then proposed the *DistilFlow* method and a tool for automatically detecting such anti-patterns and replacing them with different patterns which result in a reduction in the workflow's overall structural complexity [10] (BMC Journal paper accepted, published early 2014). This work has been performed in close collaboration with the Taverna group from the University of Manchester.

DistilFlow is part of J. Chen's thesis who has defended his PhD on October 11th, 2013 [7] and is now back to China as a research assistant in Lanzhou University.

5.5. Systems Biology

Systems Biology includes the study of interaction networks such as gene regulatory, metabolic, or signaling networks. It involves both designing the topology of the networks and predicting their dynamic and spatiotemporal aspects. It requires the import of concepts from across various disciplines and crosstalk between theory, benchwork, modelling and simulation.

5.5.1. Topological analysis of metabolic networks

In [73] we have developed a biclustering algorithm for elementary flux modes that is based on the Agglomeration of Common Motifs (ACoM). This allows a drastic diminution of the number of less significant fluxes and a kind of factorization of most important fluxes, yielding an algorithm running in quadratic time in the number of elementary flux modes.

We applied this algorithm to describe the decomposition into elementary flux modes of the central carbon metabolism in *Bacillus subtilis* and of the yeast mitochondrial energy metabolism. For *Bacillus subtilis*, a specific inhibition on the second domain of the lipoamide dehydrogenase (pdhD) component of pyruvate dehydrogenase complex that leads to the loss of all fluxes was exhibited [20]. Such a conclusion is not predictable in the classical approach.

5.5.2. Evolution of metabolic networks

A collaboration with IGM on the evolution of metabolic networks is ungoing. We aim at understanding how such networks would emerge over time among the variety of species, and how these changes could be responsible for characteristic life traits. Our methodology to characterize the evolutionary origin of the enzymatic repertoire of different fungal groups relies on machine learning. Preliminary results were presented at JOBIM 2013 [35].

5.5.3. Signaling networks

Our goal is to help the understanding of signaling pathways involving (GPCR) and to provide means to semi-automatically construct the signaling networks. Our method takes into account various kinds of biological experiments and their origin and automatically builds and draws the inferred network. Comparing the automatically deduced network with an already known fragment of the FSHR network allowed us to obtain new interesting hypotheses that are currently experimentally tested by biologists, our collaborators from INRA-BIOSin Tours. In the next months, experimental data for some GPCR (FSH, 5HT2 et 5HT4) will be prepared by BIOS and IGF (Montpellier), in the context of a GPCRNET ANR project.

Besides, in collaboration with K. Inoue, through the NII International Internship Program, we have studied the System Biology Graphical Notation language, a standard for expressing molecular networks, especially signaling networks, and proposed a translation of SBGN-AF into a logical formalism [31].

5.5.3.1. Modelling with Hsim

In a collaboration of P. Amar with microbiologists, the group of Marie-Joëlle Virolle from the *Institut de Génétique et de Microbiologie*, a first explicative model was proposed for the sigmoidicity of the shape of the survival curve of bacteria (*S. lividans*) having a antibiotic resistance gene, expressed at different levels, in presence of a constant concentration of antibiotics [24], [6], [18], [41].

This is particularly important since this method of inclusion of an antibiotics resistance gene to report the activity of its promoter is widely used in the streptomyces community.

5.5.3.2. Cancer and metabolism

It is shown in M. Behzadi's PhD thesis that most systems have very stable behaviours and that even large variations of their chemical characteristics do not affect the nature of the equilibria. This very general situation has been discovered by simulation but in some cases it is even possible to prove it mathematically.

Our collaborators M. Israël and L. Schwartz have listed more than a hundred tentative such bifurcations that we intend to study systematically. A preliminary study of the mitotic cycle with L. Paulevé has also put in evidence the strong influence of the pH of the cell on its capacity to duplicate. The PhD thesis of E. Bigan, co-directed by S. Daoudi (Univ. Denis Diderot) and J.-M. Steyaert investigates the generic properties of such complex systems and confirms that the ones we have already studied are not exceptions [43]. Some prospective cases are studied in [14].

6. Partnerships and Cooperations

6.1. Regional Initiatives

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

6.2.2. PEPS

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

6.3. European Initiatives

Program: Partenariat Hubert Curien (PHC) Procope (Jointly funded by Egide and DAAD)

Project acronym: SOSW

Project title: Sharing and Optimizing Scientific Workflows

Duration: 2013 - 2015

Coordinator: Sarah Cohen-Boulakia

International Partner

U. Humboldt (Berlin, Allemagne)

Institute for Computer Science

Ulf Leser

Abstract : Considerable effort has been put into the development of scientific workflow management systems. They support scientists in developing, running, and monitoring chains of data analysis programs. A variety of systems have reached a level of maturity that allows them to be used by scientists for their bioinformatics experiments, especially including analysis of NGS data. However, each scientific group has its own way of analyzing NGS data, using a particular set of tools, in a particular order. The aim of this project is to exploit the complementary skills of the two European groups involved to develop approaches promoting exchange of (optimized) workflows.

6.4. International Initiatives

6.4.1. Inria Associate Teams

6.4.1.1. ITSNAP

Title: Intelligent Techniques for Structure of Nucleic Acids and Proteins

Inria principal investigator: Julie Bernauer

International Partner (Institution - Laboratory - Researcher):

Stanford University (United States) - Computational Structural Biology, School of Medicine, Structural Biology - Julie Bernauer

Duration: 2012 - 2014

See also: http://www.lix.polytechnique.fr/~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.

6.4.2. Inria International Partners

6.4.2.1. Declared Inria International Partners

CARNAGE

Program: Inria-Russia

Title: CARNAGE: Combinatorics of Assembly and RNA in GEnomes

Inria principal investigator: Mireille Régnier

International Partner (Institution - Laboratory - Researcher):

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

6.4.2.2. 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.

6.4.3. Inria International Labs

R. Fonseca spent 5 months at SLAC in Stanford to work with Henry van den Bedem. J. Bernauer spent two weeks at SLAC. The associated team members also presented their work at the Inria BIS 2013 Workshop in Stanford https://project.inria.fr/inria-siliconvalley/workshops/bis2013/.

6.4.4. Participation In other International Programs

6.4.4.1. NII International Internship Program

Adrien Rougny has been an intern at NII from February to August 2013 with a support of "NII International Internship Program. He worked on the topic "Inference and Learning for Systems Biology and Network Dynamics" in Pr. Katsumi Inoue's group, a long-term collaboration of Ch. Froidevaux.

6.4.4.2. PHC Procore

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*.

6.5. International Research Visitors

6.5.1. Visits of International Scientists

H.K. Hwang

Subject: Probabilistic Analysis of A Simple Evolutionary Algorithm

Institution:Taipeh University (Taiwan)

V. Reinharz

Subject: RNA 3D structure analysis

Institution: McGill University (Canada)

E. Furletova

Subject: word enumeration

Institution: Institute of Mathematical Problems in Biology (Russia)

6.5.1.1. Internships

- C. Moutet (May and June 2013)

  Subject: Poor mappability regions in assembly

  Institution: ENS Lyon and Ecole Polytechnique Fédérale de Lausanne

  Funding: INRIA

  Supervision: M. Régnier

- F. Pirot (May and June 2013)

  Subject: Exceptional words in *Archae* genomes

  Institution: ENS Lyon

  Funding: INRIA

  Supervision: M. Régnier

- B. Fang (May to July 2013)

  Subject: Clumps combinatorics, automata and word asymptotics

  Institution: Princeton University (United States)

  Funding: Ecole Polytechnique

  Supervision: M. Régnier

- J. Moussu (April to July 2013)

  Subject: Repeats in genomic sequences

  Institution: Rennes University

  Funding: INRIA

  Supervision: M. Régnier

- M. Pichene (April to July 2013)

  Subject: Graph algorithms and protein-protein interactions

  Institution: Paris-Sud University

  Funding: INRIA

  Supervision: J. Bernauer

- L. Uroshlev (June 2013)

  Subject: Reference state for RNA KB potentials

  Institution: IOGEN (Moscou, (Russia))

  Funding: INRIA (CARNAGE)

Supervision: J. Bernauer

- O. Berillo (January and december 2013)
  Subject: miRNAs and oncogenes.
  Institution: El Farabi University (Almaty, (Kazakhstan))
  Funding: El Farabi University
  Supervision: M. Régnier
- A. Bari (March 2013)
  Subject: stress-inducible miRNAs
  Institution:El Farabi University (Almaty, (Kazakhstan))
  Funding: El Farabi University
  Supervision: M. Régnier

6.5.2. Visits to International Teams

- Sep. 2013–Sep. 2014: Y. Ponty is visiting PIMS and Simon Fraser University (Vancouver, Canada)

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 Pérès, 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. A. Denise is a member of the Scientific Committee. Y. Ponty is animator of the *Structure et interactions des macromolécules* scientific axis. C. 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: D. Saakian (A. Sinica, Taiwan), V. Reinharz (McGill), L. Tchertanov (ENS Cachan), H. Babou (Nantes), P. Ballarini (ECP), Nicolas Ferey (LIMSI), Van Du TRAN Thong (IGM), Ulf Leser (Humboldt U.), A. Zinoviev (Institut Curie, Paris), H.K. Hwang (Taipeh U.).

7.1.2.2. Other seminars

J. Bernauer presented her works at *International Conference on Biomolecular Dynamics: Experiment Meet Computation, KAUST, Saudi Arabia*.

R. Fonseca gave a talk at the Inria@SiliconValley Workshop BIS2013 in May in Stanford.

D. Iakovishina gave two talks at Institut Curie : at the weekly seminar of NGSand during the "Structural variants day" in december 2013.

7.1.2.3. International exchanges

J. Bernauer visited H. van den Bedem at SSRL (SLAC) and M. Levitt at Stanford University (USA). She visited the Huang group at HKUST (Hong-Kong).

M. Régnier and D. Iakovishina visited IoGene (Moscow).

7.1.3. Program Committee

P. Amar was a member of the steering committee and chair of the organizing committee for aSSB workshop, advances in Systems and Synthetic Biology, Nice (2013).

J. Bernauer was a member of ICBD 2013 program committee.

S. Cohen-Boulakia was member of the DILS 2013, SWEET 2013 and BDA 2013 program committees and she is member of the editorial board of the Journal on Data Semantics.

Ch. Froidevaux is a member of the editorial board of 1024, Bulletin de la Société Informatique de France, SIF.

Y. Ponty, M. Régnier, and J.-M. Steyaert served as PC members for BICOB 2013 (5th International Conference on Bioinformatics and Computational Biology, Honolulu, USA).

Y. Ponty served as PC member for ISMB/ECCB 2013 (21st International conference on Intelligent Systems for Molecular Biology/12th European Conference on Computational Biology).

M. Régnier co-organized MCCMB'13 http://mccmb.belozersky.msu.ru/2013/.

F. d'Alché-Buc, Ch. Froidevaux and Y. Ponty were members of JOBIM 2013 program committee.

J. Bernauer organized IAMB workshop (Integrative Approaches for Modeling Biomolecular Complexes) in Nice in collaboration with McGill University (Canada) and Nice University.

A one day meeting on Cancer and Metabolism was organized at LIX by J.-M. Steyaert on October 4th.

7.1.4. Research administration

- J. Bernauer is member of the IDEX Paris - Saclay Groupe de travail Sciences du Vivant.
- J. Bernauer and C. Froidevaux are member of the Comité de Pilotage of the IDEX Paris - Saclay Institut transverse de Modélisation des Sciences du Vivant.
- A. Denise is a member of the Scientific Commission of the Inria-Saclay research center. He is deputy director of the computer science department at University Paris-Sud. He is member of the Academic Senate of the Paris-Saclay University.
- Ch. Froidevaux is the head of the Bioinfo group at LRI. She was a member of a hiring committee for a Full Professor position at Polytech Paris Sud, Orsay.
- Y. Ponty is an elected member of the *Comité national du CNRS* (6th section – Foundations of Computer Science and CID 51 –Bioinformatics).
- M. Régnier is a deputy-member of DIGITEO program committee.
- J.-M.Steyaert is a member of the Board of Administrators of Polytechnique.
- J.-M. Steyaert has contributed to the organization of a workshop in July 2013 to present currently running projects between AP-HP and Polytechnique. He serves in the selection committee of a MD from HP-HP for a yearly funded research position in the Polytechnique Research Center.

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.

J.-M. Steyaert organizes BIBS (M1 and M2) at Ecole Polytechnique. Ch. Froidevaux is co-heading the Master (M1 and M2) at the University Paris Sud. 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

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

Cycle Ingénieur Polytechnicien: J. Bernauer, Algorithmes et Programmation INF421, 36h, 2ème année, Ecole Polytechnique, France

Cycle Ingénieur Polytechnicien: J. Bernauer, Modal Web Tablette INF441a, 36h, 2ème année, Ecole Polytechnique, France

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

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

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

Master : J.-M. Steyaert, X cycle ingénieur BioINF588- Algorithms for Bioinformatics, 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 : Patrick Amar, Contributions à l'étude de la dynamique des systèmes biologiques et aux systèmes de calcul en biologie synthétique, Paris Sud University, 19/12/2013

PhD : Jiuqiang Chen, Designing scientific workflows following a structure and provenance -aware strategy, Université Paris Sud, Defended on 10/10/2013, S. Cohen-Boulakia and C. Froidevaux.

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

PhD in progress : Marc Bouffard, Étude de circuits logiques moléculaires et détection de portes logiques dans un réseau métabolique, Université Paris Sud, 01/10/2013, P. Amar and F. Molina.

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: 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 : Adrien Rougny, Raisonnements sur des connaissances biologiques pour la construction et l'analyse des réseaux de signalisation, 01/10/2013, C. Froidevaux.

PhD in progress : Antoine Soulé, Evolutionary study of RNA-RNA interactions in yeast, 01/09/2013, J.-M. Steyaert 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

7.2.3. Juries

- HDR
  - Ch. Froidevaux was a reviewer for an HDR (Montpellier).
  - J.-M. Steyaert served as a jury member for Hubert Lincet HDR defence (Caen).
- PhD
  - P. Amar served as a referee for Laurent Crepin's PhD defence (Brest University).
  - Ch. Froidevaux served as a referee for a PhD thesis in Rennes and was a member of the committee for J. Leblay.
  - M. Régnier served as a referee for O. Abdou Arbi's PhD defence (Rennes University).
- Funding agencies
  - ANR 2012-2013, SIMI2, J. Bernauer and S. Cohen-Boulakia
  - UEFISCDI 2011-2013 (Research Council Romania), Y. Ponty
- Selection committees
  - Cnrs CR/DR: *comité national* (Section 6 and CID 51), Y. Ponty.
  - Inria CR2/CR1 committee: Saclay, J. Bernauer;
  - Maitre de conférence: Paris-Sud, Computer science department, S. Cohen-Boulakia and J. Azé.
  - Maitre de conférence: Bordeaux I, A. Denise.
  - Ingénieur de recherche: LIP6 UPMC (Paris), Y. Ponty.

– Chargés d'enseignement et Professeur : Ecole Polytechnique, M. Régnier et J.-M. Steyaert.

7.3. Popularization

- Outreach seminar at *Lycée Blaise Pascal* (Orsay, France) – Yann Ponty – Popular science seminar (2h), jointly organised by INRIA (Saclay) and Académie de Versailles.
- Unite ou café, Inria Saclay Popularization seminar, *Les briques de construction de la vie* , see: https://intranet.saclay.inria.fr/vie-du-centre/unithe-cafe/rencontres-2013/briques-construction-vie.

We also had the opportunity to be part of a few valorization related events. RNA structural studies were presented at the Rencontres Inria Industrie - Modélisation, simulation et calcul intensif in June 2013 http://www.inria.fr/centre/saclay/innovation/rii-modelisation-simulation-calcul-intensif/presentation. This led to an invitation at Sanofi Pharmacometry and Bioinformatics day in December 2013.

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

[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

Doctoral Dissertations and Habilitation Theses

[6] P. AMAR. , *Contributions à l'étude de la dynamique des systèmes biologiques et aux systèmes de calcul en biologie synthétique*, Université Paris Sud - Paris XI, December 2013, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00929785

[7] J. CHEN. , *Designing scientific workflows following a structure and provenance-aware strategy*, Université Paris Sud - Paris XI, October 2013, http://hal.inria.fr/tel-00931122

Articles in International Peer-Reviewed Journals

[8] O. BERILLO, A. ISSABEKOVA, M. REGNIER, A. IVASHCHENKO. *Characteristics of binding sites of intergenic, intronic and exonic miRNAs with mRNAs of oncogenes coding intronic miRNAs*, in "African Journal of Biotechnology", March 2013, vol. 12, n° 10, pp. 1016-1024 [*DOI :* 10.5897/AJB12.2054], http://hal.inria.fr/hal-00825020

[9] O. BERILLO, M. REGNIER, A. IVASHCHENKO. *Binding of intronic miRNAs to the mRNAs of host genes encoding intronic miRNAs and proteins that participate in tumourigenesis*, in "Computers in Biology and Medicine", July 2013 [*DOI :* 10.1016/J.COMPBIOMED.2013.07.011], http://hal.inria.fr/hal-00850103

[10] 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° Suppl 1, S12 p. , http://hal.inria.fr/hal-00926827

[11] A. DENISE, P. RINAUDO. *Optimisation problems for pairwise RNA sequence and structure comparison: a brief survey*, in "Transactions on Computational Collective Intelligence", 2013, to appear, http://hal.inria.fr/hal-00759573

[12] 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", 2013, vol. 10, n° 1, pp. 193-9 [*DOI :* 10.1109/TCBB.2012.148], http://hal.inria.fr/hal-00832110

[13] 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", 2013, vol. 10, n° 1, pp. 193-199, http://hal.inria.fr/hal-00756340

[14] 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, http://hal.inria.fr/hal-00933725

[15] N. LIM, Y. SENBABAOGLU, G. MICHAILIDIS, F. D'ALCHÉ-BUC. *OKVAR-Boost: a novel boosting algorithm to infer nonlinear dynamics and interactions in gene regulatory networks.*, in "Bioinformatics", June 2013, vol. 29, n° 11, pp. 1416–1423 [*DOI :* 10.1093/BIOINFORMATICS/BTT167], http://hal.inria.fr/hal-00819024

[16] 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° 12 [*DOI :* 10.1371/JOURNAL.PCBI.1003369], http://hal.inria.fr/hal-00875116

[17] A. LORENZ, Y. PONTY. *Non-redundant random generation algorithms for weighted context-free languages*, in "Theoretical Computer Science", September 2013, vol. 502, pp. 177-194 [*DOI :* 10.1016/J.TCS.2013.01.006], http://hal.inria.fr/inria-00607745

[18] V. NORRIS, P. AMAR, G. LEGENT, C. RIPOLL, M. THELLIER, J. OVADI. *Sensor potency of the moonlighting enzyme-decorated cytoskeleton*, in "BMC Biochemistry", February 2013, vol. 14, n° 3 [*DOI :* 10.1186/1471-2091-14-3], http://hal.inria.fr/hal-00766058

[19] 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

[20] S. PÉRÈS, L. FELICORI, F. MOLINA. *Elementary flux modes analysis of functional domain networks allows a better metabolic pathway interpretation*, in "PLoS ONE", 2013, http://hal.inria.fr/hal-00861577

[21] M. REGNIER, J. BOURDON. *Large deviation properties for patterns*, in "Journal of Discrete Algorithms", September 2013 [*DOI :* 10.1016/J.JDA.2013.09.004], http://hal.inria.fr/hal-00868462

[22] V. REINHARZ, Y. PONTY, J. WALDISPÜHL. *A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution.*, in "Bioinformatics", July 2013, vol. 29, n° 13, pp. i308-15, Extended version of ISMB/ECCB'13 [*DOI :* 10.1093/BIOINFORMATICS/BTT217], http://hal.inria.fr/hal-00840260

[23] 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° 11, pp. 905-19, Extended version of RECOMB'13 [*DOI :* 10.1089/CMB.2013.0085], http://hal.inria.fr/hal-00828062

[24] N. SEGHEZZI, M.-J. VIROLLE, P. AMAR. *Novel insights regarding the sigmoidal pattern of resistance to neomycin conferred by the aphII gene, in Streptomyces lividans.*, in "AMB Express", February 2013, vol. 3, n° 1, 13 p. [*DOI :* 10.1186/2191-0855-3-13], http://hal.inria.fr/hal-00794555

[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

[26] 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], http://hal.inria.fr/hal-00869141

[27] Y. ZHANG, Y. PONTY, M. BLANCHETTE, E. LECUYER, J. WALDISPÜHL. *SPARCS: a web server to analyze (un)structured regions in coding RNA sequences.*, in "Nucleic Acids Research", July 2013, vol. 41, pp. W480-5 [*DOI :* 10.1093/NAR/GKT461], http://hal.inria.fr/hal-00819017

International Conferences with Proceedings

[28] 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, http://hal.inria.fr/hal-00864645

[29] V. REINHARZ, Y. PONTY, J. WALDISPÜHL. *A linear inside-outside algorithm for correcting sequencing errors in structured RNA sequences*, in "RECOMB - 17th Annual International Conference on Research in Computational Molecular Biology - 2013", Beijing, China, 2013, http://hal.inria.fr/hal-00766781

[30] V. REINHARZ, Y. PONTY, J. WALDISPÜHL. *A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotides distribution*, in "ISMB/ECCB - 21st Annual international

conference on Intelligent Systems for Molecular Biology/12th European Conference on Computational Biology - 2013", Berlin, Germany, 2013, http://hal.inria.fr/hal-00811607

[31] A. ROUGNY, C. FROIDEVAUX, Y. YAMAMOTO, K. INOUE. *Translating the SBGN-AF language into logic to analyze signalling networks*, in "LNMR - 1st International Workshop on Learning and Non Monotonic Reasoning", La Coruña, Spain, K. INOUE, C. SAKAMA (editors), CORR, November 2013, vol. arXiv:1311.4639, pp. 44-55, http://hal.inria.fr/hal-00924230

[32] 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, China, 2013, http://hal.inria.fr/hal-00766780

[33] Y. ZHOU, Y. PONTY, S. VIALETTE, J. WALDISPÜHL, Y. ZHANG, A. DENISE. *Flexible RNA design under structure and sequence constraints using formal languages*, in "ACM-BCB - ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics - 2013", Bethesda, Washigton DC, United States, 2013, http://hal.inria.fr/hal-00823279

National Conferences with Proceedings

[34] T. BOURQUARD, D. DE VIENNE, J. AZÉ. *Identification de complexes protéine-protéine par combinaison de classifieurs. Application à Escherichia coli*, in "EGC 2013 - 13eme conférence Francophone sur l'Extraction et la Gestion des Connaissances", Toulouse, France, D. A. ZIGHED, G. VENTURINI (editors), RNTI, Hermann, January 2013, vol. E.24, pp. 419-430, http://hal.inria.fr/hal-00785473

[35] C. PEREIRA, J. AZÉ, A. DENISE, C. DREVET, C. FROIDEVAUX, P. SILAR, O. LESPINET. *Comparative analysis of phylogenetic profiles for the enzymatic characterization of fungal group*, in "JOBIM 2013", Toulouse, France, 2013, à paraître, http://hal.inria.fr/hal-00842021

Conferences without Proceedings

[36] A. FOUCHET, J.-M. DELOSME, F. D'ALCHÉ-BUC. *Gene Regulatory Network Inference using ensembles of Local Multiple Kernel Models*, in "Seventh international workshop on Machine Learning in Systems Biology, satellite meeting of ISMB'2013", Berlin, Germany, July 2013, http://hal.inria.fr/hal-00844494

[37] M. HEINONEN, O. GUIPAUD, F. MILLIAT, V. BUARD, B. MICHEAU, F. D'ALCHÉ-BUC. *Time-dependent gaussian process regression and significance analysis for sparse time-series*, in "Seventh international workshop on Machine Learning in Systems Biology, satellite meeting of ISMB'2013", Berlin, Germany, July 2013, http://hal.inria.fr/hal-00844474

[38] N. LIM, Y. SENBABAOGLU, G. MICHAILIDIS, F. D'ALCHÉ-BUC. *Boosting an operator-valued kernel-based model for gene regulatory network inference*, in "Workshop on Dynamics of biological networks: from nodes' dynamics to network evolution", Edinburgh, United Kingdom, June 2013, http://hal.inria.fr/hal-00844424

[39] N. LIM, Y. SENBABAOGLU, G. MICHAILIDIS, F. D'ALCHÉ-BUC. *Nonparametric modeling for gene regulatory network inference using boosting and operator-valued kernels*, in "Seventh International workshop on Machine Learning in Systems Biology, Satellite meeting of ISMB'2013", Berlin, Germany, July 2013, http://hal.inria.fr/hal-00844443

Scientific Books (or Scientific Book chapters)

[40] S. SCHIRMER, Y. PONTY, R. GIEGERICH. *Introduction to RNA Secondary Structure Comparison*, in "RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods", J. GORODKIN, W. L. RUZZO (editors), Methods in molecular biology, Springer, 2014, vol. 1097, http://hal.inria.fr/hal-00846818

Books or Proceedings Editing

[41] P. AMAR, F. KÉPÈS, V. NORRIS (editors). , *avances in Systems and Synthetic Biology*, EDP Sciences, March 2013, 171 p. , http://hal.inria.fr/hal-00930249

Other Publications

[42] M. ATOUI, B. BLOSSIER, V. MORÉNAS, O. PÈNE, K. PETROV. , *Semileptonic $BtoD**$ decays in Lattice QCD : a feasibility study and first results*, 2013, 28 p. , http://hal.inria.fr/hal-00917799

[43] E. BIGAN, J.-M. STEYAERT, S. DOUADY. , *Properties of Random Complex Chemical Reaction Networks and Their Relevance to Biological Toy Models*, 2013, http://hal.inria.fr/hal-00859004

[44] M. REGNIER, E. FURLETOVA, M. ROYTBERG, V. YAKOVLEV. , *Pattern occurrences Pvalues, Hidden Markov Models and Overlap Graphs*, 2014, to appear, http://hal.inria.fr/hal-00858701

References in notes

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] J. BERNAUER, S. 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

[51] 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° 3, pp. 423-425 [*DOI :* 10.1093/BIOINFORMATICS/BTR670]

[52] 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

[53] 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° 4 [*DOI :* 10.1371/JOURNAL.PONE.0018541], http://hal.inria.fr/inria-00625000

[54] 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° 4, pp. 510–528, http://dx.doi.org/10.1002/jcc.10416

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

[56] 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° 1, pp. 73-93, http://hal.inria.fr/inria-00560621

[57] 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

[58] 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

[59] 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

[60] 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

[61] A. GUILHOT-GAUDREFFROY, 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, 2014, vol. accepted

[62] 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

[63] 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

[64] 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

[65] 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

[66] 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

[67] 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

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

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

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

[71] 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° 1-2, pp. 107–127, http://www.lri.fr/~ponty/docs/Ponty-07-JMB-Boustrophedon.pdf

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

[73] S. PÉRÈS, F. VALLÉE, M. BEURTON-AIMAR, J.-P. MAZAT. *ACoM: A classification method for elementary flux modes based on motif finding*, in "BioSystems", 2011, vol. 103, n^{o} 3, pp. 410-419, http://hal.inria.fr/hal-00642137

[74] 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

[75] 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 p. [*DOI :* 10.1142/S0219720012410107], http://hal.inria.fr/hal-00757761

[76] 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

[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

[78] 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