Activity Report 2017

Team AMIBIO

Algorithms and Models for Integrative BIOlogy

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Saclay - Île-de-France

THEME
Computational Biology

Table of contents

Team AMIBIO

Creation of the Team: 2017 January 01, end of the Team: 2017 December 31

Keywords:

Computer Science and Digital Science:

A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.5. - Bayesian methods
A5.2. - Data visualization
A5.10.3. - Planning
A6.1.3. - Discrete Modeling (multi-agent, people centered)
A6.1.4. - Multiscale modeling
A6.2.3. - Probabilistic methods
A6.2.4. - Statistical methods
A6.2.6. - Optimization
A6.3.3. - Data processing
A6.3.5. - Uncertainty Quantification
A7.1.3. - Graph algorithms
A8.1. - Discrete mathematics, combinatorics
A8.2. - Optimization
A8.11. - Game Theory
A9.2. - Machine learning

Other Research Topics and Application Domains:

B1.1.1. - Structural biology
B1.1.2. - Molecular biology
B1.1.5. - Genetics
B1.1.6. - Genomics
B1.1.9. - Bioinformatics
B1.1.10. - Mathematical biology
B5.10. - Biotechnology
B9.4.1. - Computer science
B9.4.2. - Mathematics
B9.6. - Reproducibility

1. Personnel

Research Scientists

Yann Ponty [Team leader, CNRS, Researcher]
Mireille Régnier [Ecole polytechnique, Senior Researcher, HDR]

Faculty Members

Jean-Marc Steyaert [Ecole polytechnique, Professor aemeritus, HDR]

Philippe Chassignet [Ecole polytechnique, Associate Professor]

Post-Doctoral Fellow

Christelle Rovetta [Inria, from Sep 2017]

PhD Students

Alice Héliou [Ecole polytechnique, until Aug 2017]
Amélie Héliou [Ecole polytechnique, until Aug 2017]
Juraj Michalik [Inria]
Jorgelindo Moreira Da Veiga [Ecole Polytechnique/CIFRE Soredab]
Pauline Pommeret [Bourse ministérielle, since Oct 2017; Inria engineer, until Jan 2017]
Afaf Saaidi [CNRS]
Antoine Soulé [Ecole Polytechnique & Univ. McGill]
Wei Wang [Univ Paris-Sud, until Jun 2017]

Technical staff

Pauline Pommeret [Inria engineer, until Jan 2017]

Interns

Chinmay Singhal [Inria, from May 2017 until Jul 2017]
Eliott Laks [Inria, from June 2017 until Sept 2017]

Administrative Assistant

Evelyne Rayssac [Ecole polytechnique]

Visiting Scientist

Andrea Tanzer [Ecole polytechnique, Oct 2017]

2. Overall Objectives

2.1. Overall Objectives

Our project addresses a central question in bioninformatics, namely the molecular levels of organization in the cells. The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. Therefore, folding and docking are still major issues in modern structural biology and we currently concentrate our efforts on structure and interactions, and aim at a contribution towards efficient RNA design. With the recent development of computational methods 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 issues in comparative genomics.

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 close partnership with experimental biology groups. One outcome is to provide software or platform elements to predict structural models and functional hypotheses.

Increasingly, our integrative approaches have focused on problems arising in computational structural biology, with a strong focus on Bioinformatics methods focusing on the sequence(s) to structure(s) relationship in RiboNucleic Acids (RNAs). RNAs are versatile biomolecules found in all domains of life, of length ranging from 20-30 nucleotides (nts) in micro RNAs to dozens of thousands nucleotides in certain messenger RNAs or coronaviruses. In most functional families, the structure (or lack thereof) adopted by an RNA is instrumental to its mission(s), and is the object of considerable identifiable pressure throughout evolution. Understanding the structure of RNA and its dynamics leads to testable functional hypotheses. Conversely, a deeper understanding of how function requires the adoption of certain structures leads to models and tools for the rational design of RNAs. AMIBio develops methods and algorithms to predict the dynamics of folding, to make sense of low-dimensional experimental data, perform a rational design of functional RNAs, and detect instances of RNA families within genomic and transcriptomic data.

3. Research Program

3.1. RNA and protein structures

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.

3.1.1. Discrete representations and complexity

Participants: Yann Ponty, Wei Wang, Antoine Soulé, Juraj Michalik.

*Common activity with J. Waldispühl (McGill) and A. Denise (*LRI*).*

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 [46], the correction of pyrosequencing errors [38], and the efficient detection of multi-stable RNAs (riboswitches) [42], [43].

Increasingly, we develop and study parameterized complexity approaches, based on dynamic programming over a tree decomposition, for several combinatorial problems, including RNA design, structure-sequence alignment (aka threading in the context of proteins). The later problem is at the core of Wei Wang's Phd, successfully defended in Dec 2017. In the context of our probabilistic approaches, often based on random generation, such parameterized algorithms usually follow proofs of hardness for the associated enumeration problems.

3.1.2. RNA design.

Participants: Alice Héliou, Yann Ponty.

Joint project with A. Denise (sc Lri), J. Waldispühl (McGill), D. Barash (Univ. Ben-Gurion), and C. Chauve (Simon Fraser University).

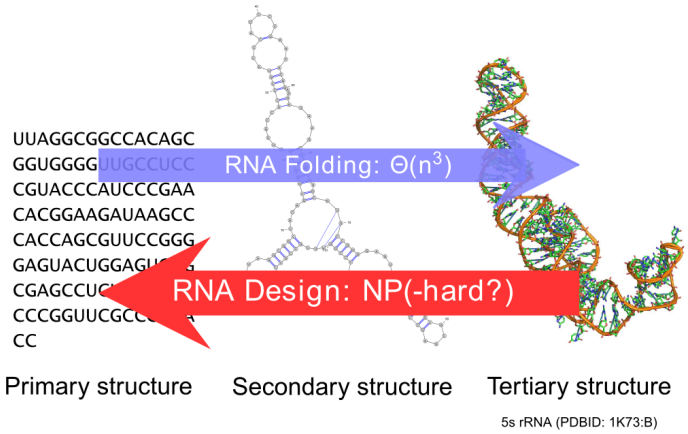

Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

It is a natural pursuit 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 AMIBio, we develop a new methodology, based on weighted random generation [25] 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 [32].

The main advantages of this approach is its linear complexity, and its flexibility in incorporating constraints. Indeed, extensive experiments revealed a drift of existing software towards sequences of high G+C-content, and we showed how to control this distributional bias by using multidimensional Boltzmann sampling [37], [36]. Recently, we are extending this approach to the design of RNAs with multiple structures, developping a Fixed-Parameter Tractable framework that naturally extends to capture negative design goals.

3.1.3. Modeling large macromolecular architectures

Participants: Yann Ponty, Afaf Saaidi, Mireille Régnier, Amélie Héliou.

Joint projects with A. Denise (LRI), D. Barth (Versailles), J. Cohen (Paris-Sud), B. Sargueil (Paris V) and Jérome Waldispühl (McGill).

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 contribute methods inspired by both of these directions.

We also develop homology-based approaches for structure modeling, and developed a general setting for the problem of RNA structure-sequence alignment, known to be NP-hard in the presence of complex topological features named pseudoknots (PKs). Our approach is based on tree decomposition of structures and gives rises to a general parameterized algorithm, where the exponential part of the complexity depends on the family of structures [39]. This work unifies and generalizes a number of recent works on specific families, and enables the curation of multiple alignments for RNA families featuring PKs, correcting certain bias introduced by PK-oblivious methods.

3.2. Séquences

Participants: Mireille Régnier, Philippe Chassignet, Yann Ponty, Jean-Marc Steyaert, Alice Héliou, 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. Among specific motifs involved in molecular interactions, one may cite protein-DNA (cis-regulation), protein-protein (docking), RNA-RNA (miRNA, frameshift, circularisation). 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. A collaboration has beenestablished with LOB, at Ecole Polytechnique, who bought a sequencing machine, through the co-advised thesis of Alice Héliou. 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 [34]. 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. Combinatorial Algorithms and motifs

Participants: Mireille Régnier, Philippe Chassignet, Alice Héliou.

Besides applications [41] of analytic combinatorics to computational biology problems, the team addressed general combinatorial problems on words and fundamental issues on languages and data structures. Motif detection combines an algorithmic search of potential sites and a significance assessment. To assess the significance of an observation usually requires the evaluation of a quantitative criterion such as the P-value. In the recent years, a general scheme of derivation of analytic formula for the P-value under different constraints (k-occurrence, first occurrence, overrepresentation in large sequences,...) has been provided. It relies on a representation of continuous sequences of overlapping words, currently named *clumps* or *clusters* in a graph [35]. Recursive equations to compute p-values may be reduced to a traversal of that graph, leading to a linear algorithm. This improves over the space and time complexity of the generating function approach or previous probabilistic weighted automata.

In [45], it is claimed that half of the genome consists of different types of repeats. One may cite microsatellites, DNA transposons, transposons, long terminal repeats (LTR), long interspersed nuclear elements (LINE), ribosomal DNA, short interspersed nuclear elements (SINE). Therefore, knowledge about the length of repeats is a key issue in several genomic problems, notably assembly or re-sequencing. Preliminary theoretical results are given in [29], and, recently, heuristics have been proposed and implemented [26], [40], [23]. A dual problem is the length of minimal absent words. Minimal absent words are words that do not occur but whose proper factors all occur in the sequence. Their computation is extremly related to finding maximal repeats (repeat that can not be extended on the right nor on the left). The comparison of the sets of minimal absent words provides a fast alternative for measuring approximation in sequence comparison [22], [24].

Recently, it was shown that considering the words which occur in one sequence but do no in another can be used to detect biologically significant events [44]. We have studied the computation of minimal absent words and we have provided new linear implementations [20]. We are now working on a dynamic approach to compute minimal absent words for a sliding window. For a sequence of size n, we expect a complexity of $O(n)$ time and space, independent of the size of the window. This approach could be use to align a sequence on a larger sequence using minimal absent words for comparison.

3.2.2. Random generation

Participants: Yann Ponty, Juraj Michalik, Christelle Rovetta.

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 [27]: 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.

We also introduced algorithms and data structures for a non-redundant generation of combinatorial objects. In situtations where the search space of a problem can be unambigously explored using dynamic programming, such algorithms generate objects within a postulated distribution, conditioned to avoid previously generated objects. This method can be used to probe objects having lower probabilities, a desirable property in the context of RNA kinetics studies, or could lead to better estimators in context where the exact emission probability of each object can be computed.

3.3. 3D interaction and structure prediction

Participant: Amélie Héliou.

The biological function of macromolecules such as proteins and nucleic acids relies on their dynamic structural nature and their ability to interact with many different partners. This is specially challenging as structure flexibility is key and multi-scale modelling [21], [28] and efficient code are essential [33].

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.

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. At Inria, our interest for structural biology is shared by the ABS and ORPAILLEUR project-teams. 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.

Game theory was used by M. Boudard in her PhD thesis, defended in 2015, to predict the 3d structure of RNA. In her PhD thesis, co-advised by J. Cohen (LRI), A. Héliou extended the approach to the prediction of protein structures.

3.3.1. Robotics-inspired structure and dynamics

Participant: Amélie Héliou.

We recently work one a robotics approach to sample the conformational space of macromolecules like RNAs [1]. The robotics approach allows maintaining the secondary structure of the RNA fixed, as an unfolding is very unlikely and energetically demanding. By this approach we also dramatically reduce the number of degrees of freedom in the molecule. The conformational space becomes possible to be sampled. This reduction does not reduce the quality of the sampling.

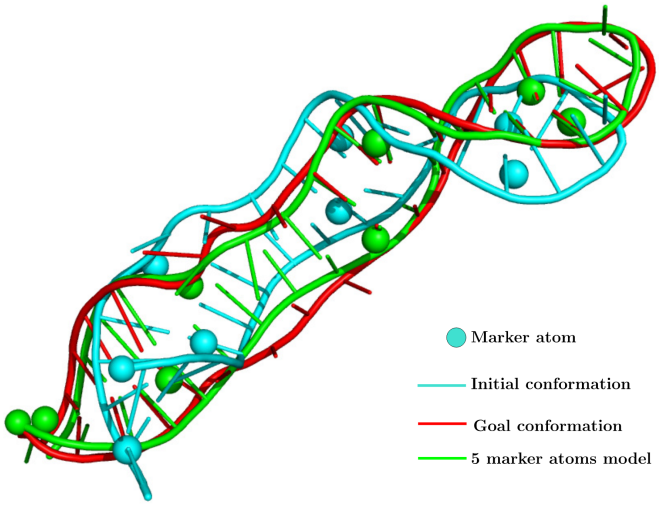

Figure 2. The cyan structure is the initial conformation, the red structure is the goal conformation. The full-atom initial conformation was driven toward the goal conformation using only the position of the goal sphere atoms. The green conformation is the result obtained; spheres perfectly overlap with the goal position and the overall conformation is really close to the goal conformation.

Our current work consists in applying the same approach to a targeted move. The motion is then driven either by the position of a few atoms or the distances between couple of atoms. Theses two aspects are under development and will increase our capacity to integrate experimental data. Our method can drive a RNA conformation toward another conformation of the same RNA given only the position of a few atoms (marker atoms).

For instance double electron-electron resonance (DEER) experimental results are distributions of distances. Probes are attached to the molecules and the distances between to probes is measured and outputted as a distribution. Our method is able to sample an ensemble of all-atom conformations that can explain the distance distribution.

3.3.2. Game theory and molecular folding

Participant: Amélie Héliou.

Building on a previous collaboration with LRI (J. Cohen, expert in algorithmic game theory – GALAC team), we have extended an original approach to structure prediction based on game theory, previously introduced as a proof-of-concept [31]. Our model of the folding model as a game represents an alternative paradigm to thermodynamics and energy minimization, in which individual residues are considers as selfish players, whose strategy (position/orientation) is driven by a desire to increase their local utility function (consistency

with observed local conformations). The conformations where residues can no longer gain any utility by changing their local conformations are called Nash equilibria. They can be thought of as local minima of a suitably-defined energy function, and can sometimes be efficiently computed using stochastic strategies.

Our work is first to find an algorithm that can guarantee the convergence to an Nash equilibrium (a state were no player would increase his payoff by playing something different alone) and prove their convergence. At the same time, we are looking for efficient and biologically relevant ways of defining the game settings so that Nash equilibria correspond to folded states. One direction would be to draw a parallel between Nash equilibria and local minima of the kinetic landscape. This line of research also raises questions related to learning Nash equilibria.

4. Application Domains

4.1. Circular RNAs

Participants: Mireille Régnier, Alice Héliou.

Circular RNAs (circRNAs) have been found abundantly in human cells as well as in many other animals. These non-coding RNAs are involved in the regulation of numerous biological processes, and it was recently shown that, as pre-miRNA, they might may actually encode short functional peptides. Our collaborators at Ecole Polytechnique (Biology Dept, LOB) have demonstrated the role of RNA ligase *Pab1020* in RNA circularization. The protein *Pab1020* is a member of the conserved *Rnl3* family of RNA ligases that are predominantly found in hyperthermophiles (archaea, bacteria) and halophiles.

Many computational methods have been proposed to identify and characterize circular RNA from high throughput sequencing data. However, they all suffer from a low specificity, leading to an explosion of false positives. Along with our partners at LOB (Ecole Polytechnique), we develop a robust method for the detection of circRNAs, particularly well-suited to accomodate to analyze sequencing data acquired in extreme environments.

4.2. Analysis of probing data

Participants: Yann Ponty, Mireille Régnier, Afaf Saaidi.

SHAPE probing [47] is an experimental technique in which RNA is exposed to a reagent which, upon reverse-transcription, induces a modification (truncation, mutation) in the DNA. The prevalence of such modifications, which depends on the locally adopted structure(s) (or lack thereof), can be measured for each nucleotide using sequencing techniques, informing regarding the 2D structure. SHAPE probing data can thus be used by structure prediction methods, either to assess their consistency with a proposed structural model, or to restrict the conformation space.

As part of a colloboration with B. Sargueil's lab (Faculté de pharmacie, Paris V) funded by the Fondation pour la Recherche medical, we strive to propose a new paradigm for the analysis data produced using a new experimental technique, called SHAPE analysis (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension). This experimental setup produces an accessibility profile associated with the different positions of an RNA, the *shadow* of an RNA. We currently design new algorithmic strategies to infer the secondary structure of RNA from multiple SHAPE experiments performed by experimentalists at Paris V. Those are obtained on mutants, and will be coupled with a fragment-based 3D modeling strategy developed by our partners at McGill.

5. New Results

5.1. New circular RNAs identified in Pyrococcus abyssi

We contributed a new method for the detection of circRNAs, which we validated on simulated data, and used to analyze the transcriptome of *Pyrococcus abyssi*, an archae living at high depth and temperature [1]. Using this method, which was shown to produce less false positives than previous computational approaches, we analyzed data produced in collaboration with LOB (Ecole Polytechnique), and detected roughly a hundred of novel candidates circular RNAs. Moreover, we provided evidence, on a large scale, that the protein *Pab1020* acts as a ligase, and interacts with some of these circular RNAs, shedding new light on the mechanisms underlying the circularization process.

5.2. Minimal absent words

Minimal absent words are words that do not occur but whose proper factors all occur in the sequence. In a collaboration with King's College, several algorithms, we have designed algorithms to search for minimal absent words in external memory [8], and *in-line*, using a sliding window [13] (parallelization, external memory,...) that outperform previous solutions and achieve near-optimal speed up. This opens new scenarios in the applications of minimal absent words in computational biology, including phylogeny or evolution. For instance, it was shown that there exist three minimal words in Ebola virus genomes which are absent from human genome. As two strings coincide iff they have the same set of minimal absent words, an interesting side result is to solve in optimal time the pattern matching problem using *negative information*.

5.3. Kinematics-inspired algorithms for macromolecular modeling

At a geometric level, RNA is much more flexible than protein, and undergoes smooth transitions between its various conformations. Such transitions are difficult to observe, but can be predicted using algorithms inspired by kinematics and motion-planning. With our partners at Stanford, we designed and implemented such an algorithm within the KGS library [8] to morph between two RNA conformations while keeping distance constraints induced by base pairs and, more importantly, avoiding clashes. In a more preliminary work, we also used similar approaches to automagically fit multi-conformer ligand models into electron density maps [16].

5.4. RNA design

In a paper published in Algorithmica [6], we have shown that our previous results [30] hold for more sophisticated energy models where base-pairs are associated with arbitrary energy contributions. This result, which required a complete overhaul of our previous proofs (e.g. using arguments based on graph coloring), allows us to foresee an extension of (at least some of) our results to state-of-the-art models, such as the Turner energy model.

In collaboration with Danny Barash's group at Ben-Gurion university (Israel), we contributed a review of existing tools and techniques for RNA design, which was published in *Briefings in Bioinformatics* series [3].

Finally, in a paper [14] recently accepted for a presentation at the prestigious RECOMB'18 conference, we revisited the problem of generating at random an RNA sequence which is simultaneously compatible with a set of target secondary structures. This problem was previously addressed by our collaborators at the TBI Vienna/Univ. Leipzig, using an exponential-time algorithm. We established the #*P*-hardness of the problem, and its inapproximibility in general. However, the problem is still amenable to an efficient parametrization, and we proposes an FTP algorithm named `RNARedPrint` based on the tree decomposition for the random generation, to which we adapted a multidimensionnal Boltzmann sampling technique in order to gain (probabilistic) control over secondary features such as the *GC%*, the relative free-energy of the various structures...

5.5. Game theory and macromolecular modeling

Initially based on a very coarse representation of RNA, we refined our model of RNA folding as a game, using on-lattice coordinates and statistical potentials for the utility function. The resulting algorithm was implemented in the subsequent version of the `GARN` [2] software.

The final year of Amélie Héliou's PhD led to theoretical developments in game theory, mainly obtained in collaboration with J. Cohen (LRI, Univ. Paris-Sud). First, the quasi-exponential convergence, under reasonable assumptions, of the HEDGE algorithm was demonstrated in collaboration with the POLARIS team in Grenoble [11]. Moreover, in a paper accepted at NIPS'17 [12], we addressed the learning of Nash equilibria. In this context, we established the convergence with high probability of no-regret learning in the bandit and semi-bandit settings.

5.6. RNA kinetics using non-redundant sampling

RNA kinetics is arguably the next frontier in RNA 2D bioinformatics. In particular, computational methods for studying the kinetics of RNA beyond 150nts are hindered by the combinatorial explosion of the conformation space. In an effort to circumvent such an effect, we have proposed a sampling approach that explicitly target local minima of the energy function. Our sampling algorithm, jointly proposed with H. Touzet (Bonsai, Inria Lille & CrisTaL, Univ. Lille I) and accepted for a presentation at the ISMB/ECCB'17 conference in Prague [15], uses non-redundant sampling principles to avoid an excessive concentration of samples within low local minima.

5.7. New insight from SHAPE probing data

Existing computational methods for structure prediction are typically hindered by their assumption of a single structure, and their assumption of orthogonal signals stemming from different reagents. To overcome these limitations, we contributed an integrative approach combining stochastic sampling and structural clustering [17] (journal version pending). In collaboration with ENS Lyon/Univ. Lyon I and Univ. Paris-Descartes, we used this method to model the structure of the HIV-1 gag open-reading frame [4].

We also addressed the problem of binning sets of NGS reads arising from the simultaneous probing, using the SHAPEmap protocol, of variants produced by a error-prone PCR. We proposed a variant of the Expectation-Maximization algorithm [10] to jointly infer maximum-likelihood origins for reads and mutational profiles for each variant.

6. Partnerships and Cooperations

6.1. National Initiatives

6.1.1. FRM

AMIBio is in charge of Bioinformatics developments in this project on structural prediction from RNA probing data (SHAPE). It involves Biochemists at Université Paris Descartes (France, PI B. Sargueil) and is funded by a "Fondation pour la Recherche Medicale" grant. It also involves partners in Paris-Sud (France) and McGill University (Canada).

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

- Approche comparatives haut-débit pour la modelisation de l'architecture 3D des ARN à partir de données experimentales
- 2015–2018
- Yann Ponty, A. Denise, M. Regnier, A. Saaidi (PhD funded by FRM)
- B. Sargueil (Paris V – Experimental partner), J. Waldispuhl (Univ. McGill)

6.2. European Initiatives

6.2.1. Collaborations in European Programs, Except FP7 & H2020

Yann Ponty is the French PI for the French/Austrian RNALANDS project, jointly funded by the French ANR and the Austrian FWF, in partnership with the Theoretical Biochemistry Institute (University of Vienna, Austria), LRI (Univ. Paris-Sud) and EPI BONSAI (Inria Lille-Nord Europe).

French/Austrian International Program

RNALANDS (ANR-14-CE34-0011)

Fast and efficient sampling of structures in RNA folding landscapes

01/10/2014–30/09/2018

Coordinated by AMIB (Inria Saclay) and TBI Vienna (University of Vienna)

EPI BONSAI/INRIA Lille - Nord Europe, Vienna University (Austria), LRI, Université Paris-Sud (France)

The main goal of the RNALands project is to provide efficient tools for studying the kinetics of RiboNucleic Acids, based on efficient sampling strategies.

6.3. International Initiatives

6.3.1. Inria Associate Teams Not Involved in an Inria International Labs

6.3.1.1. ALARNA

Title: Associated Laboratory for the Analysis of RiboNucleic Acids

International Partner (Institution - Laboratory - Researcher):

McGill University (Canada) - REUSSI Program - Jerome Waldispuhl

Start year: 2017

See also: https://team.inria.fr/alarna/

RiboNucleic Acids (RNAs) are ubiquitous biomolecules whose structure, adopted as the outcome of a complex folding process, often plays a crucial part in cellular processes. The ALARNA Associate Team (Laboratory for the Analysis of RiboNucleic Acids), which consist of the AMIBio project-team (Inria Saclay/Ecole Polytechnique, France) and the CSB (Computer Science and Biology) group at university McGill (Montreal, Canada), addresses key questions in RNA bioinformatics. More specifically, it dedicates much of its effort to the production and interpretation of chemical probing data generated by SHAPE, an experimental technology which allows to accurately predict, in a high-throughput, one or several secondary structure(s) adopted by an RNA. To that end, the teams contribute their unique combinations of expertise, ranging from combinatorial optimization to sequence algorithmics through structural bioinformatics.

6.3.1.2. Informal International Partners

AMIBio enjoys regular interactions with the following institutions:

- TBI, University of Vienna (Austria). Within the RNALands project funded by the Austrian FWF and the french ANR, we frequently interact with our partners at the TBI, on projects associated with the kinetics of RNAs. Over the course of 2017, we have visited our partners twice, once in Vienna and once in Bled (Slovenia) over the course of the 2017 Winter retreat of the TBI. Additionally, Andrea Tanzer has visited AMIBio for a month in Oct 2018, funded by a visiting scholar program of Ecole Polytechnique;
- Simon Fraser University (Vancouver, Canada). The Mathematics department at SFU has ongoing projects on RNA design, comparative genomics and RNA structure comparison with our team. M. Mishna (SFU) has visited Inria Saclay in January 2017 to push an ongoing collaboration on 2D walks;

- McGill University (Montréal, Canada). Following our productive collaboration with J. Waldispühl (Computer Science Dept, McGill), and the recent defense of V. Reinharz's PhD, whose thesis was co-supervised by AMIBio members, we have increased our interactions on SHAPE data analysis through the ALARNA associate team;
- King's college (London, UK). Our collaboration with L. Mouchard (AMIBio associate) and S. Pissis on string processing and data structures was at the core of Alice Héliou's PhD, defended in July 2017.

6.3.2. Participation in Other International Programs

Title: PHC GRO-algo – Combination of time-course GRO-seq assay, algorithmics and software development for measuring genome-wide transcription elongation rates

International Partner (Institution - Laboratory - Researcher):

Wuhan University (China), College of Life Science – Pr Yu Zhou

Start year: 2017

Participant in a French-Chinese Hubert Curien Partnerships (PHC), supported by CampusFrance and funding bilateral exploratory research exchanges in Bioinformatics. The program involves research scientists from Wuhan University, Ecole Polytechnique and Univ. Paris-Sud.

Title: Computational methods and databases to identify small RNA-binding molecules regulating gene expression

International Partner (Institution - Laboratory - Researcher):

University McGill (Canada), Computer Science & Biochemistry – J. Waldispühl, N. Moitessier; Univ. Strasbourg, IBMC - E. Westhof.

Start year: 2017

The project, headed by N. Moitessier and J. Waldispühl (McGill University, Canada) strives to develop tools to derive a mechanical understanding of riboswitches at the 2D and 3D levels, including chemoinformatics aspects.

6.4. International Research Visitors

6.4.1. Visits of International Scientists

Andrea Tanzer

Date: Oct 2017 - Nov 2017

Institution: TBI Vienna, Austria

Mathieu Blanchette

Date: June 2017

Institution: Univ. McGill, Canada

6.4.1.1. Internships

Paul Arijit

Institution: IISc Bangalore (India)

Supervisor: Mireille Régnier

Chinmay Singhal

Date: May 2017 - July 2017

Institution: IIT Guwahati, India (India)

Supervisor: Yann Ponty

7. Dissemination

7.1. Promoting Scientific Activities

7.1.1. Scientific Events Organisation

7.1.1.1. General Chair, Scientific Chair

- RECOMB'18: Y. Ponty, M. Regnier (co-chair)

7.1.2. Scientific Events Selection

7.1.2.1. Member of the Conference Program Committees

- RECOMB'17: Y. Ponty, M. Regnier
- ISMB/ECCB'17: Y. Ponty
- ACM/BCB'17: A. Héliou
- MCCMB'17: M. Régnier
- BiCoB'17: Y. Ponty

7.1.2.2. Reviewer

- CPM'17: Y. Ponty

7.1.3. Journal

7.1.3.1. Member of the Editorial Boards

M. Régnier is an editor of PeerJ Computer Science.

7.1.3.2. Reviewer - Reviewing Activities

M. Régnier and Y. Ponty reviewed manuscripts for a large selection of journals in Mathematics, Computer Science and Bioinformatics: Discrete Mathematics and Theoretical Computer Science, Theoretical Computer Science, Bioinformatics, BMC Bioinformatics, Journal of Mathematical Biology, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Journal of Discrete Algorithms, Algorithms for Molecular Biology, PLOS One, Journal of Theoretical Biology, RNA, Nucleic Acids Research...

7.1.4. Leadership within the Scientific Community

Y. Ponty is *animateur* of the *Macromolecular structure and interactions* axis of the CNRS GDR BIM (BioInformatique Moléculaire). With F. Cazals (ABS, Inria Sophia-Antipolis), he co-created in Oct. 2017 and currently heads the MASIM (Méthodes Algorithmiques pour les Structures et Interactions des Macromolécules) workgroup of BIM;

M. Régnier is a member of DIGITEO progam Committee and SDV working group in Saclay area.

7.1.5. Scientific Expertise

Y. Ponty reviewed grants for the OPUS program of the NCN agency (Poland), the Bioinformatics and Theoretical Biology programm of the DFG (Germany). He acted as an external reviewer and external assessor for an assistant professor position at the Faculty of Science of the University of South Denmark. He also refereed for a postdoc call by the French Charmmmat LabEx of Univ. Paris-Saclay.

7.1.6. Research Administration

M. Regnier is the current head of the Computer Science department (LIX) of Ecole Polytechnique.

7.2. Teaching - Supervision - Juries

7.2.1. Teaching

7.2.1.1. Initial training in engineering at Ecole Polytechnique.

At the undergraduate level, AMIBio is essentially involved in Computer Science courses at Ecole Polytechnique, mainly in programming languages and Big Data processing (high-performance computing and machine learning). Notably, we are involved in the *parcours d'Approfondissement en Bioinformatique* at École Polytechnique.

7.2.1.2. Graduate-level programs and courses.

Our project is also very much involved in the AMI2B (*Analyse, Modélisation et Ingénierie de l'Information Biologique et Médicale*, formerly named BIBS – *Bioinformatique et Biostatistique*) Master program at Université Paris-Sud/École Polytechnique. Most AMIBio permanent members teach recurrent courses in AMI2B, and M. Regnier is in charge of the program at the M1 and M2 levels for Ecole Polytechnique.

Beyond the *plateau de Saclay*, AMIBio members participate in the BIM master program at UPMC, and regularly deliver PhD-level lectures in Summer/Winter schools.

Bachelor (Licence & Ecole Ingénieur)

- P. Chassignet, INF*, 100h, L2/L3, Ecole Polytechnique, France
- Al. Héliou, INF*, 40h, L3, Ecole Polytechnique, France

Master

- P. Chassignet, INF*/Modal Bioinfo, 90h, M1/M2, Ecole Polytechnique, France
- Al. Héliou, INF*, 7h, M1 AMI2B, Univ. Paris-Saclay, France
- Am. Héliou, INF*, 44h, M1 AMI2B, Univ. Paris-Saclay, France
- Y. Ponty, Bioinfo ARN, 12h AMI2B, M2, Univ. Paris-Saclay, France
- Y. Ponty, Bioinfo ARN, 12h BIM, M2, UPMC, France
- M. Regnier, Algo/Combinatoire, 12h AMI2B, M2, Univ. Paris-Saclay, France
- J.-M. Steyaert, Algo/Combinatoire, 12h AMI2B, M2, Univ. Paris-Saclay, France
- J.-M. Steyaert, Bioinformatics INF/BIO 588, 20h Majeure BioInfo, M1, Ecole Polytechnique,France

Doctorat

- Y. Ponty, Design ARN, 3h BIM, PhD, Univ. Tehran, Iran

7.2.2. Supervision

PhD :

Wei Wang, *Practical structure-sequence alignment of pseudoknotted RNAs*, Univ. Paris-Saclay, December 2017, Y. Ponty, A. Denise

Alice Héliou, *Analyse des séquences génomiques : Identification des ARNs circulaires et calcul de l'information négative*, Univ. Paris-Saclay, July 2017, M. Regnier

Amélie Héliou, *Théorie des jeux et échantillonnage de conformations pour l'amarrage macromoléculaire multi-corps et multi-échelle*, Univ. Paris-Saclay, August 2017, J. Cohen

PhD in progress

Afaf Saaidi, *Differential analysis of RNA SHAPE probing data*, Ecole Polytechnique, Encadrants: Yann Ponty and Mireille Régnier.

Antoine Soulé, *Evolutionary study of RNA-RNA interactions in yeast* , Ecole Polytechnique, Encadrants: Jean-Marc Steyaert and J. Waldispuhl (U. McGill, Canada);

Jorgelindo Moreira da Veiga, *Caractérisation dynamique et optimisation des flux métaboliques* , Ecole Polytechnique, Encadrants: L. Schwartz (AP-HP)Sabine Peres (U. Paris-Sud)

Juraj Michalik, *Non-redundant sampling for the study of RNA kinetics*, Inria Saclay, Encadrants: Y. Ponty and H. Touzet (Cristal, Univ. Lille I);

Pauline Pommeret, *Étude de l'impact phénotypique de variants génétiques altérant la structure secondaire des ARNs chez le bovin : Méthodes algorithmiques et validations expérimentales*, INRA Jouy & Ecole Polytechnique, Encadrants: Y. Ponty and D. Rocha (INRA Jouy);

Ha Thi Ngoc Nguyen, *New computational strategies for analyzing the diversity of transcriptome sequence and structure, and their relationship to disease*, I2BC (Univ. Paris-Sud) & Ecole Polytechnique, Encadrants: Y. Ponty and D. Gautheret (Univ. Paris-Sud);

7.2.3. Juries

Y. Ponty participated in a hiring committee for an associate professor position in Discrete Mathematics at the University of Southern Denmark;

Y. Ponty acted as an external opponent for the PhD defence of X. Pan at the University of Copenhagen;

7.3. Popularization

May 2017 Participation to the CURIOSITas art & science festival organized by univ. Paris-Saclay on the CNRS campus (Gif-sur-Yvette, France);

March 2017 Outreach presentation by AMIBio on RNA folding and combinatorial design for a class of high school students (Montmorency, France);

8. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] H. F. BECKER, A. HÉLIOU, K. DJAOUT, R. LESTINI, M. REGNIER, H. MYLLYKALLIO. *High-Throughput Sequencing Reveals Circular Substrates for an Archaeal RNA ligase*, in "RNA Biology", March 2017, https://hal.archives-ouvertes.fr/hal-01491132

[2] M. BOUDARD, D. BARTH, J. BERNAUER, A. DENISE, J. COHEN. *GARN2: coarse-grained prediction of 3D structure of large RNA molecules by regret minimization*, in "Bioinformatics", 2017, vol. 16, pp. 2479-2486 [*DOI :* 10.1093/BIOINFORMATICS/BTX175], https://hal.archives-ouvertes.fr/hal-01589347

[3] A. CHURKIN, M. D. RETWITZER, V. REINHARZ, Y. PONTY, J. WALDISPÜHL, D. BARASH. *Design of RNAs: comparing programs for inverse RNA folding*, in "Briefings in Bioinformatics", January 2017 [*DOI :* 10.1093/BIB/BBW120], https://hal.inria.fr/hal-01392958

[4] J. DEFORGES, S. DE BREYNE, M. AMEUR, N. ULRYCK, N. CHAMOND, A. SAAIDI, Y. PONTY, T. OHLMANN, B. SARGUEIL. *Two ribosome recruitment sites direct multiple translation events within HIV1 Gag open reading frame*, in "Nucleic Acids Research", July 2017, vol. 45, n° 12, pp. 7382–7400 [*DOI :* 10.1093/NAR/GKX303], https://hal.archives-ouvertes.fr/hal-01505282

[5] W. DUCHEMIN, Y. ANSELMETTI, M. PATTERSON, Y. PONTY, S. BÉRARD, C. CHAUVE, C. SCORNAVACCA, V. DAUBIN, E. TANNIER. *DeCoSTAR: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies*, in "Genome Biology and Evolution", 2017, vol. 9, n° 5, pp. 1312-1319, https://hal.inria.fr/hal-01503766

[6] J. HALEŠ, A. HÉLIOU, J. MAŇUCH, Y. PONTY, L. STACHO. *Combinatorial RNA Design: Designability and Structure-Approximating Algorithm in Watson-Crick and Nussinov-Jacobson Energy Models*, in "Algorithmica", November 2017, vol. 79, n° 3, pp. 835–856, https://arxiv.org/abs/1603.03577 [*DOI :* 10.1007/S00453-016-0196-X], https://hal.inria.fr/hal-01285499

[7] A. HÉLIOU, D. BUDDAY, R. FONSECA, H. VAN DEN BEDEM. *Fast, clash-free RNA conformational morphing using molecular junctions*, in "Bioinformatics", July 2017, vol. 33, n° 14, pp. 2114 - 2122 [*DOI :* 10.1093/BIOINFORMATICS/BTX127], https://hal.archives-ouvertes.fr/hal-01569620

[8] A. HÉLIOU, S. P. PISSIS, S. J. PUGLISI. *emMAW: Computing Minimal Absent Words in External Memory*, in "Bioinformatics", 2017 [*DOI :* 10.1093/BIOINFORMATICS/BTX209], https://hal.archives-ouvertes.fr/hal-01569271

[9] B. LÖWES, C. CHAUVE, Y. PONTY, R. GIEGERICH. *The BRaliBase dent - a tale of benchmark design and interpretation*, in "Briefings in Bioinformatics", March 2017, vol. 18, n° 2, pp. 306–311 [*DOI :* 10.1093/BIB/BBW022], https://hal.inria.fr/hal-01273406

Invited Conferences

[10] A. SAAIDI, Y. PONTY, M. BLANCHETTE, M. REGNIER, B. SARGUEIL. *An EM algorithm for mapping short reads in multiple RNA structure probing experiments*, in "Matbio2017", London, United Kingdom, King's College London , September 2017, https://hal.inria.fr/hal-01590528

International Conferences with Proceedings

[11] J. COHEN, A. HÉLIOU, P. MERTIKOPOULOS. *Hedging under uncertainty: regret minimization meets exponentially fast convergence*, in "Symposium on Algorithmic Game Theory (SAGT) 2017", L'Aquila, Italy, Proceedings of the 10th International Symposium on Algorithmic Game Theory, September 2017, https://arxiv.org/abs/1607.08863 [*DOI :* 10.1007/978-3-319-66700-3_20], https://hal.archives-ouvertes.fr/hal-01382290

[12] J. COHEN, A. HÉLIOU, P. MERTIKOPOULOS. *Learning with bandit feedback in potential games*, in "NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems", Long Beach, CA, United States, December 2017, https://hal.archives-ouvertes.fr/hal-01643352

[13] M. CROCHEMORE, A. HÉLIOU, G. KUCHEROV, L. MOUCHARD, S. P. PISSIS, Y. RAMUSAT. *Minimal absent words in a sliding window & applications to on-line pattern matching*, in "FCT 2017", Bordeaux, France, Lecture Notes in Computer Science, Springer, September 2017, forthcoming, https://hal.archives-ouvertes.fr/hal-01569264

[14] S. HAMMER, Y. PONTY, W. WANG, S. WILL. *Fixed-Parameter Tractable Sampling for RNA Design with Multiple Target Structures*, in "RECOMB 2018 – 22nd Annual International Conference on Research in Computational Molecular Biology", Paris, France, April 2018, https://hal.inria.fr/hal-01631277

[15] J. MICHÁLIK, H. TOUZET, Y. PONTY. *Efficient approximations of RNA kinetics landscape using non-redundant sampling*, in "ISMB/ECCB - 25th Annual international conference on Intelligent Systems for Molecular Biology/16th European Conference on Computational Biology - 2017", Prague, Czech Republic, July 2017, vol. 33, n° 14, pp. i283 - i292 [*DOI :* 10.1093/BIOINFORMATICS/BTX269], https://hal.inria.fr/hal-01500115

[16] G. V. ZUNDERT, D. KEEDY, P. SURESH, A. HÉLIOU, K. BORRELLI, T. DAY, J. FRASER, H. VAN DEN BEDEM. *Objectively and automatically building multi-conformer ligand models in electron densities*, in "Conformational ensembles from experimental data and computer simulations", Berlin, Germany, August 2017, https://hal.inria.fr/hal-01569829

Conferences without Proceedings

[17] A. SAAIDI, Y. PONTY, B. SARGUEIL. *An integrative approach for predicting the RNA secondary structure for the HIV–1 Gag UTR using probing data*, in "JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques", Lille, France, July 2017, 1 p. , https://hal.archives-ouvertes.fr/hal-01534587

Other Publications

[18] C. CHAUVE, J. COURTIEL, Y. PONTY. *Counting, generating, analyzing and sampling tree alignments*, 2017, Submitted to IJFCS, https://hal.inria.fr/hal-01500116

[19] D. SURUJON, Y. PONTY, P. CLOTE. *Small-world networks and RNA secondary structures*, January 2017, working paper or preprint, https://hal.inria.fr/hal-01424452

References in notes

[20] C. BARTON, A. HELIOU, L. MOUCHARD, S. PISSIS. *Linear-time computation of minimal absent words using suffix array*, in "BMC Bioinformatics", 2014, vol. 15, 11 p. [*DOI :* 10.1186/s12859-014-0388-9], https://hal.inria.fr/hal-01110274

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

[22] S. CHAIRUNGSEE, M. CROCHEMORE. *Using minimal absent words to build phylogeny*, in "Theoretical Computer Science", 2012, vol. 450, n^{o} 0, pp. 109-116

[23] R. CHIKHI, P. MEDVEDEV. *Informed and automated k-mer size selection for genome assembly*, in "Bioinformatics", Jan 2014, vol. 30, n^{o} 1, pp. 31–37, http://dx.doi.org/10.1093/bioinformatics/btt310

[24] M. CROCHEMORE, G. FICI, R. MERCAS, S. PISSIS. *Linear-Time Sequence Comparison Using Minimal Absent Words*, in "LATIN 2016: Theoretical Informatics - 12th Latin American Symposium", Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2016, http://arxiv.org/abs/1506.04917

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

[26] H. DEVILLERS, S. SCHBATH. *Separating significant matches from spurious matches in DNA sequences*, in "Journal of Computational Biology", 2012, vol. 19, n^{o} 1, pp. 1–12, http://dx.doi.org/10.1089/cmb.2011.0070

[27] 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

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

[29] Z. GU, H. WANG, A. NEKRUTENKO, W. H. LI. *Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence*, in "Gene", Dec 2000, vol. 259, n^{o} 1-2, pp. 81–88

[30] J. HALEŠ, J. MAŇUCH, Y. PONTY, L. STACHO. *Combinatorial RNA Design: Designability and Structure-Approximating Algorithm*, in "Annual Symposium on Combinatorial Pattern Matching", Springer, 2015, pp. 231–246

[31] 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^{o} 1, pp. 193-199, http://hal.inria.fr/hal-00756340

[32] 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^{o} 20, pp. 10041-52 [*DOI :* 10.1093/NAR/GKS768], http://hal.inria.fr/hal-00733924

[33] 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^{o} 8, pp. 1127-8 [*DOI :* 10.1093/BIOINFORMATICS/BTQ083], http://hal.inria.fr/inria-00536404

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

[35] M. REGNIER, E. FURLETOVA, M. ROYTBERG, V. YAKOVLEV. *Pattern occurrences Pvalues, Hidden Markov Models and Overlap Graphs*, 2013, submitted, http://hal.inria.fr/hal-00858701

[36] 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^{o} 13, i308 p. [*DOI :* 10.1093/BIOINFORMATICS/BTT217], http://hal.inria.fr/hal-00840260

[37] 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, Allemagne, 2013, http://hal.inria.fr/hal-00811607

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

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

[40] G. RIZK, D. LAVENIER, R. CHIKHI. *DSK: k-mer counting with very low memory usage*, in "Bioinformatics", Mar 2013, vol. 29, n^{o} 5, pp. 652–653 [*DOI :* 10.1093/BIOINFORMATICS/BTT020], http://bioinformatics.oxfordjournals.org/content/early/2013/02/01/bioinformatics.btt020.full

[41] 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

[42] 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

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

[44] R. M. SILVA, D. PRATAS, L. CASTRO, A. J. PINHO, P. J. S. G. FERREIRA. *Three minimal sequences found in Ebola virus genomes and absent from human DNA*, in "Bioinformatics", 2015 [*DOI :* 10.1093/BIOINFORMATICS/BTV189]

[45] T. J. TREANGEN, S. L. SALZBERG. *Repetitive DNA and next-generation sequencing: computational challenges and solutions*, in "Nat Rev Genet", Jan 2012, vol. 13, n^{o} 1, pp. 36–46, http://dx.doi.org/10.1038/nrg3117

[46] 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

[47] K. A. WILKINSON, E. J. MERINO, K. M. WEEKS. *Selective 2 [prime]-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution*, in "Nature protocols", 2006, vol. 1, n^{o} 3, pp. 1610–1616