IN PARTNERSHIP WITH:
CNRS

Ecole Polytechnique

Université Paris-Sud (Paris 11)

Activity Report 2016

Project-Team AMIB

Algorithms and Models for Integrative Biology

IN COLLABORATION WITH: Laboratoire d'informatique de l'école polytechnique (LIX)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Computational Biology

Table of contents

Project-Team AMIB

Creation of the Team: 2009 May 01, updated into Project-Team: 2011 January 01, end of the Project-Team: 2016 December 31

Keywords:

Computer Science and Digital Science:

3.4. - Machine learning and statistics
3.4.5. - Bayesian methods
7. - Fundamental Algorithmics
7.2. - Discrete mathematics, combinatorics

Other Research Topics and Application Domains:

1. - Life sciences
1.1. - Biology
1.1.1. - Structural biology
1.1.9. - Bioinformatics
1.1.10. - Mathematical biology
9.6. - Reproducibility

1. Members

Research Scientists

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

Faculty Members

Philippe Chassignet [Ecole Polytechnique, Associate Professor]
Laurent Mouchard [Université de Rouen, Associate Professor, HDR]
Jean-Marc Steyaert [Ecole Polytechnique, Professor, HDR]

Engineer

Pauline Pommeret [Inria]

PhD Students

Alice Héliou [Ecole Polytechnique]
Amélie Héliou [Ecole Polytechnique]
Juraj Michalik [Inria]
Jorgelindo Moreira Da Veiga [CIFRE Soredab]
Afaf Saaidi [CNRS]
Antoine Soulé [Ecole Polytechnique, until Oct 2016]
Wei Wang [Université Paris-Sud]

Administrative Assistant

Evelyne Rayssac [Ecole Polytechnique]

2. Overall Objectives

2.1. Overall Objectives

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

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. Dynamic programming 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 [44], the correction of pyrosequencing errors [37], and the efficient detection of multi-stable RNAs (riboswitches) [40], [41].

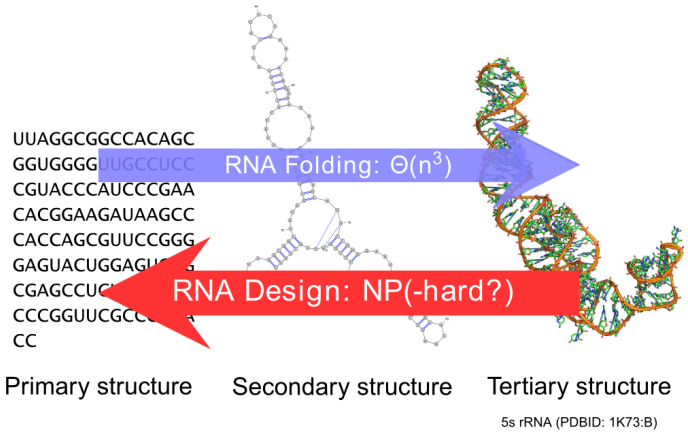


Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

3.1.2. RNA design.

Participants: 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).

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 [24] 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 [31].

3.1.3. Towards 3D modeling of large molecules

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 settings directions.

Modeling tasks can also be greatly helped by the availability of experimental data. However, high-resolution techniques such as crystallography or RMN, are notoriously costly in term of time and ressources, leading to the current gap between the amount of available sequences and structural data. 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. As part of A. Saaidi's PhD, 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.

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 [33]. 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 [39] 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. Assessment significance requires a quantitative criterion such as the p-value.

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 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.

This research area is widened by new problems arising from *de novo* genome assembly or re-assembly.

In [43], 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 [28], and, recently, heuristics have been proposed and implemented [25], [38], [22]. 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 [21], [23].

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 [42]. We have studied the computation of minimal absent words and we have provided new linear implementations [18],[16]. 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) in 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.

According to the current knowledge, cancer develops as a result of the mutational process of the genomic DNA. In addition to point mutations, cancer genomes often accumulate a significant number of chromosomal rearrangements also called structural variants (SVs). Identifying exact positions and types of these variants may lead to track cancer development or select the most appropriate treatment for the patient. Next Generation Sequencing opens the way to the study of structural variants in the genome, as recently described in [20]. This is the subject of an international collaboration with V. Makeev's lab (IOGENE, Moscow), MAGNOME project-team and V. Boeva (Curie Institute). One goal is to combine two detection techniquesbased either on paired-end mapping abnormalities or on variation of the depth of coverage. A second goal is to develop a model of errors, including a statistical model, that takes into account the quality of data from the different sequencing technologies, their volume and their specificities such as the GC-content or the mappability.

3.2.2. Random generation

Participants: Yann Ponty, Juraj Michalik.

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 [26]: 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, in a joint work with A. Denise (LRI), we have done significant and original progress in this area recently [34], [39], 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 [33].

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 [19], [27] and efficient code are essential [32].

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 is extending the approach to predict 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.

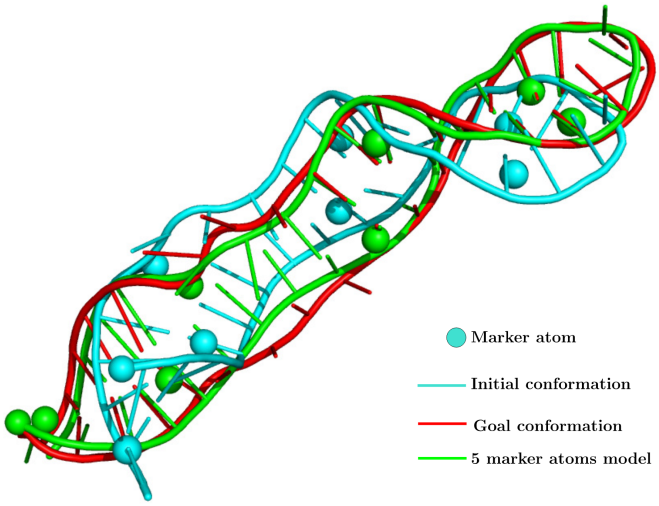


Figure 2. The cyan structure is the initial conformation, the red structure is the goal conformation. The full-atom initial conformation was driven toward the goal conformation using only the position of the goal sphere atoms. The green conformation is the result obtained; spheres perfectly overlap with 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 the analysis possibility of 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 protein folding

Participant: Amélie Héliou.

M.Boudard used game theory to sample folded conformations of RNA. We work in apply game theory to sample folded conformations of proteins. This is challenging as a protein is generally less flexible than a RNA and thus accept less conformations.

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.

4. New Software and Platforms

4.1. GenRGenS

GENeration of Random GENomic Sequences
KEYWORDS: Bioinformatics - Genomic sequence
FUNCTIONAL DESCRIPTION

A software dedicated to the random generation of sequences. Supports different lasses of models, including weighted context-free grammars, Markov models, ProSITE patterns..

- Participants: Yann Ponty and Alain Denise
- Contact: Yann Ponty
- URL: https://www.lri.fr/~genrgens/

4.2. VARNA

Interactive drawing and editing of the RNA secondary structure
KEYWORDS: Bioinformatics - Structural Biology
SCIENTIFIC DESCRIPTION

VARNA is Java lightweight Applet dedicated to drawing the secondary structure of RNA. It is also a Swing component that can be very easily included in an existing Java code working with RNA secondary structure to provide a fast and interactive visualization.

Being free of fancy external library dependency and/or network access, the VARNA Applet can be used as a base for a standalone applet. It looks reasonably good and scales up or down nicely to adapt to the space available on a web page, thanks to the anti-aliasing drawing primitives of Swing.
FUNCTIONAL DESCRIPTION

Varna is a new tool for the automated drawing, visualization and annotation of the secondary structure of RNA, designed as a companion software for web servers and databases.

Varna implements four drawing algorithms, supports input/output using the classic formats dbn, ct, bpseq and RNAML and exports the drawing as five picture formats, either pixel-based (JPEG, PNG) or vector-based (SVG, EPS and XFIG).

It also allows manual modification and structural annotation of the resulting drawing using either an interactive point and click approach, within a web server or through command-line arguments.

- Participants: Alain Denise and Yann Ponty
- Partners: CNRS - Ecole Polytechnique - Université Paris-Sud
- Contact: Yann Ponty
- URL: http://varna.lri.fr/

5. New Results

5.1. RNA design

We extended our previous results on RNA design [29], obtained in collaboration with J. Hales, J. Manuch and L. Stacho (Simon Fraser University/Univ. British Columbia, Canada).

Our results provided complete characterizations for the structures that can be designed using restricted alphabets. We provided a complete characterization of designable structures without unpaired bases. When unpaired bases are allowed, we provided partial characterizations for classes of designable/undesignable structures, and showed that the class of designable structures is closed under the stutter operation. Membership of a given structure to any of the classes can be tested in linear time and, for positive instances, a solution could be found in linear time. Finally, we considered a structure-approximating version of the problem that allows to extend helices and, assuming that the input structure avoids two motifs, we provided a linear-time algorithm that produces a designable structure with at most twice more base pairs than the input structure, as illustrated by Fig. 3.

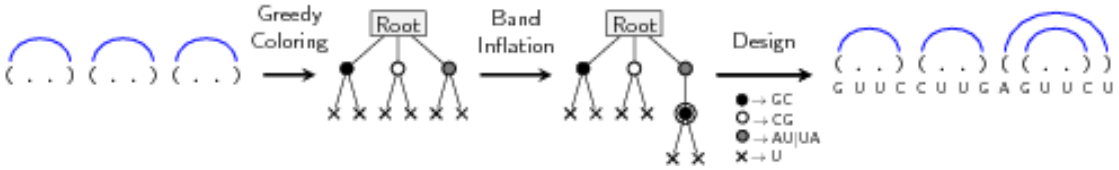


Figure 3. Principle of our structure-approximating version of RNA design: Starting from a potentially undesignable structure, a greedy coloring can be performed and corrected such that the final structure is provably designable in linear time.

In a manuscript accepted for publication in Algorithmica [4], we have shown that our previous results [29] 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.

We also initiated a collaboration with Danny Barash's group at Ben-Gurion university (Israel). We contributed a review of existing tools and techniques for RNA design, to appear as an article within the *Briefings in Bioinformatics* series [2]. We also combined previously contributed methods for design into a new method and web-server for the design of RNAs [3]. This collaboration stemed from the observation that `IncaRNAtion` [36], a random generation algorithm for RNA design recently developed in collaboration with Jérôme Waldispühl's group at McGill University (Montreal, Canada), produced excellent starting points (seed) for classic algorithms based on local-search. In particular, the combination of `IncaRNAtion` and `RNAfbInv` [45] was found to yield particularly promising candidates for design.

5.2. Algorithmics and combinatorics of motifs occurrences

We have developed a new algorithm to compute minimal absent words in external memory. Minimal absent words are used in sequence comparison [23] or to detect biologically significant events. For instance, it was shown that there exist three minimal words in Ebola virus genomes which are absent from human genome [42]. The identification of such specific-species sequences may prove to be useful for the development of diagnosis and therapeutics. We have already provide an $O(n)$-time and $O(n)$-space algorithm to compute minimal absent words, with an implementation that can be executed in parallel. However these implementations require a large amount of RAM, thus they cannot be used for the human genome on a desktop computer. In our new contribution we developed an external memory implementation, it can compute minimal absent words of length at most 11 for the human genome using only 1GB of RAM in less than 4 hours (manuscript submitted [16]).

Repetitive patterns in genomic sequences have a great biological significance. This is a key issue in several genomic problems as many repetitive structures can be found in genomes. One may cite microsatellites, retrotransposons, DNA transposons, long terminal repeats (LTR), long interspersed nuclear elements (LINE), ribosomal DNA, short interspersed nuclear elements (SINE). Knowledge about the length of a maximal repeat also has algorithmic implications, most notably the design of assembly algorithms that rely upon de Bruijn graphs.

Analytic combinatorics allowed us to derive formula for the expected length of repetitions in a random sequence [9]. The originality of the approach is the demonstration of a Large Deviation principle and the use of Lagrange multipliers. This allowed for a generalization of previous works on a binary alphabet. Simulations on random sequences confirmed the accuracy of our results. As an application, the sample case of Archaea genomes illustrated how biological sequences may differ from random sequences, and in turns provides tools to extract repetitive sequences.

5.3. Integrative RNA structural modeling

To circumvent expensive, low-throughput, 3D experimental techniques such as X-ray crystallography, a low resolution/high throughput technology called SHAPE is increasingly favored for structural modeling by structural biologists.

Within Afaf Saaidi's thesis, funded by the *Fondation pour la Recherche Médicale* and co-supervised by Bruno Sargueil at Faculté de Pharmacie of Université Paris V, we have developed integrative modeling strategies based on Boltzmann sampling. Preliminary results, obtained by applying these methods to model the structures of 3'UTR regions in Ebola, were presented at JOBIM 2016 [14].

Moreover, in collaboration with McGill University (Canada), we cross-examined mutate-and-map data (MaM [30]) in the light of evolutionary data. MaM data consist in the sequential SHAPE probing of a set of mutant RNAs, obtained by systematic point-wise mutations, to highlight structurally-dependent nucleotides, later to use dependent pairs as constraints in (an automated) structural modeling. We chose to adopt an alternative perspective on MaM data, and used the perturbation of the SHAPE profiles as a proxy for the structural disruption induced by a mutation. Disruptive mutations are rescued within homologs, *i.e.* compensated to re-establish the structure. However, our analysis also revealed the existence of non-structurally local (neither on the 2D or 3D levels) nucleotides which have significant mutual-information with highly disruptive positions, despite not being involved in any obvious compensatory relationship.

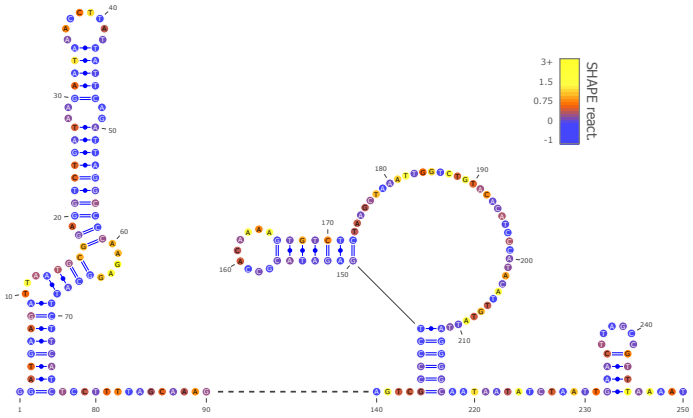
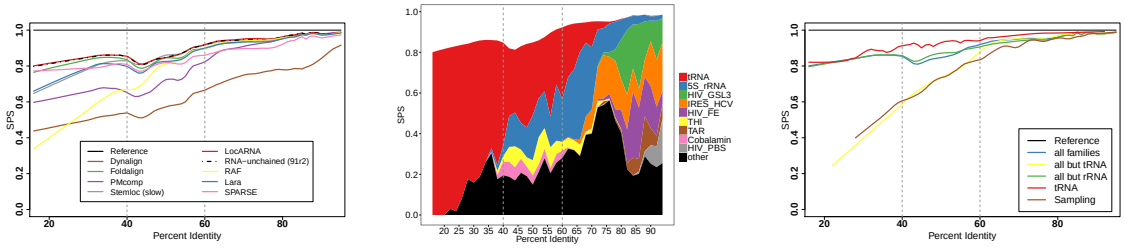


Figure 4. Conserved and thermodynamically-stable structure elements revealed by our analysis of an Ebola UTR region.

We hypothesized that such mutations are revealing of interactions involving RNA. In a manuscript published in *Nucleic Acids* journal, we tested and validated this hypothesis by showing its capacity to discriminate discriminative positions that are known to be in contact with specific ligands (proteins, DNA, small molecules...) [10].



a. *b.* *c.*

Figure 5. Typical software for comparative RNA structure prediction exhibit a dent in performance within the 40%-60% sequence identity range when benchmarked using the popular Bralibase data set (a.). However, this is due to the overrepresentation of a well-predicted type of RNA (tRNAs, red area) for low-identity ranges (b.). A re-evaluation of state-of-the-art software on an unbiased (c., brown line) reveals much more modest predictive capacities than initially believed in the community.

A fruitful line of research for RNA structure prediction is based on a comparative approach. Whenever homologous RNAs are identified, a classic strategy is to perform a simultaneous alignment and folding of several RNAs. Many software (30+) have been contributed over the past decades for this problem, leading to the introduction of benchmarks, one of the most prominent being the `Bralibase`, to position new developments and identify axes of progression. One such desired improvement, as illustrated in Figure 5, was the difficulties experienced by most software around the 40-60% sequence identity range, which was believed to arise from deep algorithmic reasons. In collaboration with Cedric Chauve (Simon Fraser University, Canada) Benedikt Löwes and Robert Giegerich (Bielefeld University, Germany), we showed that this perceived difficulty was simply

the consequence of a strong bias towards tRNAs in the 40-60% sequence identity region. Moreover, we argued that the overall performance of existing tools for low sequence identities were largely overestimated [8].

Finally, we presented at JOBIM 2016 an efficient implementation, called `LicoRNA`, of our parameterized complexity algorithm based on tree-decomposition for the sequence/structure alignment of RNA [15]. Specifically, we showed that our `LicoRNA`, by including an expressive scoring scheme and capturing pseudoknots of arbitrary complexity, generally outperforms previously contributions for the problem.

5.4. Combinatorial foundations

Pairwise ordered tree alignment are combinatorial objects that appear in RNA secondary structure comparison. However, the usual representation of tree alignments as supertrees is ambiguous, i.e. two distinct supertrees may induce identical sets of matches between identical pairs of trees. This ambiguity is uninformative, and detrimental to any probabilistic analysis. In a recent collaboration with Cédric Chauve (SFU Vancouver, Canada) and Julien Courtiel (LIPN, Paris XII) presented at the ALCOB'16 conference, we considered tree alignments up to equivalence [11]. Our first result was a precise asymptotic enumeration of tree alignments, obtained from a context-free grammar by means of basic analytic combinatorics. Our second result focused on alignments between two given ordered trees. By refining our grammar to align specific trees, we obtained a decomposition scheme for the space of alignments, and used it to design an efficient dynamic programming algorithm for sampling alignments under the Gibbs-Boltzmann probability distribution. This generalizes existing tree alignment algorithms, and opens the door for a probabilistic analysis of the space of suboptimal RNA secondary structures alignments.

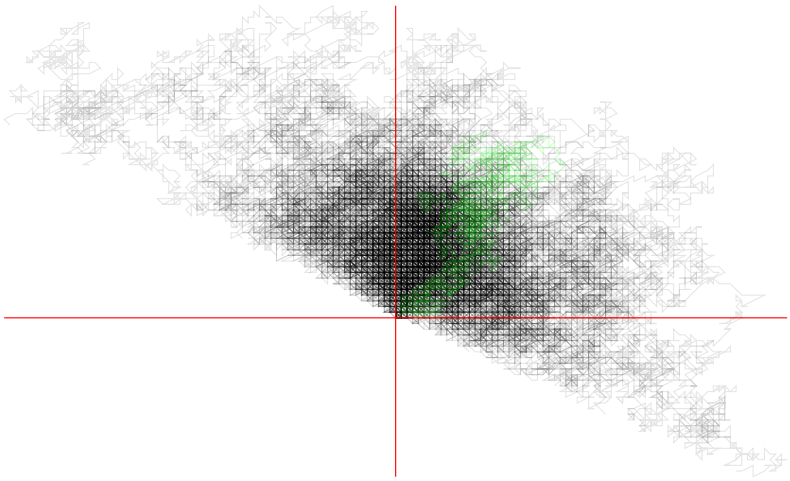


Figure 6. Random 2D walks (green walks) confined in the positive can be generated efficiently by performing rejection from a well-chosen 1D model (black walks) [12].

Finally, in collaboration with Marni Mishna (Simon Fraser University, Canada) and Jérémie Lumbroso (Princeton University, USA), we considered the uniform random generation of *difficult*, or *reluctant*, 2D discrete walks that remain confined in the positive quarter plane. We proposed a naive dynamic programming algorithm having complexity $O(n^4)$ for any step set. We also exploited the remark that any quarterplane walks can be transformed into a well-chosen 1D model having the same exponential growth factor. However, such a 1D model takes irrational steps, leading us to explore new avenues for the random generation. This work was presented at the GASCOM'16 conference [12].

5.5. Comparative genomics

D. Iakovishina defended in 2015 a PhD thesis co-advised by M. Régnier and V. Boeva (Curie Institute). She proposed a new computational method to detect structural variants using whole genome sequencing data. It combines two techniques that are based either on the detection of paired-end mapping abnormalities or on the detection of the depth of coverage. SV-BAY relies on a probabilistic Bayesian approach and includes a modelization of possible sequencing errors, read mappability profile along the genome and changes in the GC-content. Keeping only somatic SVs is an additional option when matched normal control data are provided. SV-BAY compares favorably with existing tools on simulated and experimental data sets [6] Software SV-BAY is freely available https://github.com/InstitutCurie/SV-Bay.

As a side product, a novel exhaustive catalogue of SV types -to date the most comprehensive SV classification- was built. On the grounds of previous publications and experimental data, seven new SV types, ignored by the existing SV calling algorithms, were exhibited.

We also contributed, in collaboration with Céline Scornavacca's group (ISEM, Montpellier) to the algorithmic foundations of the `EcceTERA` software [7] for the reconciliation of gene and species phylogenetic trees. This software adopts a maximum parsimony approach to predict in an evolutionary model that includes duplications, losses and transfers of genes.

6. Partnerships and Cooperations

6.1. National Initiatives

6.1.1. FRM

Yann Ponty is the Bioinformatics PI for a *Fondation de la Recherche Médicale*-funded project.

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.

Partenariat Henri Curien (CampusFrance, programme Staël)

Random constrained permutations: Algorithms and Analysis

01/01/2015–31/12/2016

Coordinated by CMAP (Ecole Polytechnique) and Maths Dept (Univ. Zürich, Switzerland)

LIX (Ecole Polytechnique), LIPN (Univ Paris XIII), LIGM (Univ Marne-la-Vallée)

The goals of this collaborative network is to initiate or push several collaborations related to the structure of random constrained permutations. AMIB bring their strength in random generation and discrete algorithms, and benefit from the considerable expertise accumulated at University of Zürich on permutation patterns.

6.3. International Initiatives

6.3.1. Inria International Labs

6.3.1.1. Declared Inria International Partners

Title: AMAVI - Combinatorics and Algorithms for the Genomic sequences

International Partner (Institution - Laboratory - Researcher):

Vavilov Institute of General Genetics (Russia (Russian Federation)) - Department of Computational Biology - Vsevolod Makeev

Duration: 2013 - 2017

Start year: 2013

VIGG and AMIB teams has a more than 12 years long collaboration on sequence analysis. The two groups aim at identifying DNA motifs for a functional annotation, with a special focus on conserved regulatory regions. In the current 3-years project CARNAGE, our collaboration, that includes Inria-team MAGNOME, is oriented towards new trends that arise from Next Generation Sequencing data. Combinatorial issues in genome assembly are addressed. RNA structure and interactions are also studied.

The toolkit is pattern matching algorithms and analytic combinatorics, leading to common software.

6.3.1.2. Regular International Partners

AMIB enjoys regular interactions with the following institutions:

- 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) will also visit Inria Saclay in January 2017 to push an oingoing 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 AMIB members, we plan to increase our interactions on SHAPE data analysis by applying for an Inria associate team;
- King's college (London, UK). Our collaboration with L. Mouchard (AMIB associate) and S. Pissis on string processing and data structures is at the core of Alice Héliou's PhD. To finalize the implementation of her algorithms and apply them on real data, Alice has spent a two month period during the summer of 2016 at the EBI.

6.3.2. Participation in Other International Programs

France-Stanford exchange program

Duration: 2014 - 2016

Start year: 2014

See also: http://francestanford.stanford.edu/collaborative_projects

Amélie Héliou is co-supervised by H. Van Den Bedem in Stanford. Her tow-months visit to Stanford during the Fall of 2016 was funded by France-Stanford.

6.4. International Research Visitors

6.4.1. Visits of International Scientists

6.4.1.1. Internships

Frédéric Lavner

Date: 01/07/2016- 31/08/2016

Institution: ENSEA (France)

Supervisor: Mireille Régnier

Maria Waldl

Date: 01/08/2016 - 31/09/2016

Institution: TBI, University of Vienna (Austria)

Supervisor: Yann Ponty

6.4.2. Visits to International Teams

- Yann Ponty has visited M. Mishna and C. Chauve at the Simon Fraser University for two weeks in July 2016;
- Juraj Michalik has visited A. Tanzer and I. Hofacker at the university of Vienna (Austria) for one week in November 2016;
- Mireille Régnier has visited MIPT (Moscow) and Novossibirsk University to enhance student exchanges between these universities and Ecole polytechnique.

6.4.2.1. Research Stays Abroad

- Alice Héliou has visited the EBI (UK) for two months during the Fall of 2016;
- Amélie Héliou has visited Stanford University (USA) for two months during the Summer of 2016;

7. Dissemination

7.1. Promoting Scientific Activities

7.1.1. Scientific Events Selection

7.1.1.1. Member of the Conference Program Committees

- Yann Ponty: RECOMB'17, BICOB'17, SeqBio'16, ECCB'16, BioVis'16, ISMB'16, and BICOB'16

7.1.1.2. Reviewer

For international conferences:

- Yann Ponty: MFCS'16

7.1.2. Journal

7.1.2.1. Member of the Editorial Boards

M. Régnier is an editor of PeerJ Computer Science.

7.1.2.2. 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...

7.1.3. Invited Talks

Mireille Regnier was an invited speaker of the *Advanced Algorithms on Strings* workshop in the honor of Alberto Apostolico.
Yann Ponty gave invited talks at the *Journées Combinatoires Franco-Vancouvéroises* (Vancouver, Canada), the college of life sciences (Wuhan University, China), and multiple seminars (C3BI@Pasteur, LIGM, LAMSADE, 2x Bioinfo@LIX/LRI...)

7.1.4. Leadership within the Scientific Community

Yann Ponty is the scientific animator of the *macromolecular structure and interactions axis* of the CNRS *Molecular Bioinformatics* workgroup (GdR BIM).

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

7.1.5. Scientific Expertise

Yann Ponty acted as an external expert for the French *Agence Nationale de la Recherche* (ANR, JCJC program), and for the Canadian *Sciences and Engineering Research Council* (NSERC/CRSNG, Discovery grant program);

Yann Ponty acted as an external reviewer for the evaluation of the assistant professor position of Jing Qin at University of Southern Denmark, towards her promotion as an associate professor;

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

7.1.6. Research Administration

Since 2016, M. Regnier acts as the head of LIX (CNRS/Ecole Polytechnique);

Until Sept. 2016, Yann Ponty was an elected member of the *comité national du CNRS*, and took part in the evaluation of CNRS research scientists and structures at a national level in Computer Science (Section 6) and Life Science interfaces (CID 51);

Yann Ponty is an elected member of the *conseil de laboratoire* of LIX.

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 AMI2B at Paris-Saclay (previously BIBS (Bioinformatique et Biostatistique) at Université Paris-Sud/École Polytechnique) and the parcours d'Approfondissement en Bioinformatique at École Polytechnique.

At Ecole Polytechnique, M. Régnier is in charge of M1 and M2. Most team members are teaching in this master program.

Beyond the plateau de Saclay, Yann Ponty taught 12h at the M2 level for University Pierre et Marie Curie in the BIM Master program.

7.2.2. Supervision

HDR

L. Mouchard *Contributions algorithmiques à l'analyse des séquences génomiques* : J.-M. Steyaert

PhD

Vladimir Reinharz, *Algorithmic properties of evolved structured RNAs*, McGill University (Montréal, Canada), July 2016, supervised by J. Waldispühl (CS Dept, McGill Univ.) and Yann Ponty

PhD in progress

Alice Héliou, *Identification et caractérisation d'ARN circulaires dans des séquences NGS*, Ecole Polytechnique, Encadrant(e|s): Mireille Régnier and Hubert Becker

Wei Wang, *Homology based approaches for predicting 3D structure of RNA molecules*, Univ. Paris XI, Encadrant(e|s): Alain Denise and Yann Ponty;

Amélie Héliou, *Game theory and conformation sampling for multi-scale and multi-body macromolecule docking*, Ecole Polytechnique, Encadrant(e|s): Johanne Cohen;

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)

7.2.3. Juries

HDR

S. Bérard *Histoires évolutives et autres comptes* : M. Régnier

M. Magnin *Contributions à l'élaboration de connaissances qualitatives en bio-informatique* : M. Régnier

PhD

Manuel Lafond, Comparative Genomics, Université de Montréal, Canada, August 2016

Karen Druart, Computational Structural Biology, Ecole Polytechnique, December 2016

7.3. Popularization

Afaf Saaidi participated to the *Ma thèse en trois minute* contest, and won the best poster award at the *Journées de l'école doctoral Interfaces* of Paris Saclay University.

8. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] E. BIGAN, L. PAULEVÉ, J.-M. STEYAERT, S. S. DOUADY. *Necessary and sufficient conditions for protocell growth*, in "Journal of Mathematical Biology", April 2016 [*DOI :* 10.1007/s00285-016-0998-0], https://hal.archives-ouvertes.fr/hal-01338156

[2] A. CHURKIN, M. DRORY RETWITZER, V. REINHARZ, Y. PONTY, J. WALDISPÜHL, D. BARASH. *Design of RNAs: Comparing Programs for inverse RNA folding*, in "Briefings in Bioinformatics", 2017, forthcoming, https://hal.inria.fr/hal-01392958

[3] M. DRORY RETWITZER, V. REINHARZ, Y. PONTY, J. WALDISPÜHL, D. BARASH. *incaRNAfbinv : a web server for the fragment-based design of RNA sequences*, in "Nucleic Acids Research", 2016, vol. 44, n° W1, pp. W308 - W314 [*DOI :* 10.1093/NAR/GKW440], https://hal.inria.fr/hal-01319682

[4] 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", 2016, forthcoming, https://hal.inria.fr/hal-01285499

[5] A. HELIOU, M. LÉONARD, L. MOUCHARD, M. SALSON. *Efficient dynamic range minimum query*, in "Theoretical Computer Science", 2017 [*DOI :* 10.1016/J.TCS.2016.07.002], https://hal.archives-ouvertes.fr/hal-01255499

[6] D. IAKOVISHINA, I. JANOUEIX-LEROSEY, E. BARILLOT, M. REGNIER, V. BOEVA. *SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read map-pability*, in "Bioinformatics (Oxford, England)", January 2016, https://hal.inria.fr/hal-01253126

[7] E. JACOX, C. CHAUVE, G. J. SZÖLLŐSI, Y. PONTY, C. SCORNAVACCA. *ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony*, in "Bioinformatics (Oxford, England)", July 2016, pp. 2056-8, Accepted, forthcoming, https://hal.inria.fr/hal-01276903

[8] B. LÖWES, C. CHAUVE, Y. PONTY, R. GIEGERICH. *The BRaliBase dent-a tale of benchmark design and interpretation.* , in "Briefings in Bioinformatics", March 2016 [*DOI :* 10.1093/BIB/BBW022], https://hal.inria.fr/hal-01273406

[9] M. REGNIER, P. CHASSIGNET. *Accurate prediction of the statistics of repetitions in random sequences: a case study in Archaea genomes*, in "Frontiers in Bioengineering and Biotechnology", May 2016, https://hal.inria.fr/hal-01304366

[10] V. REINHARZ, Y. PONTY, J. WALDISPÜHL. *Combining structure probing data on RNA mutants with evolutionary information reveals RNA-binding interfaces*, in "Nucleic Acids Research", 2016, vol. 44, n° 11, e104 - e104 [*DOI :* 10.1093/NAR/GKW217], https://hal.inria.fr/hal-01291754

International Conferences with Proceedings

[11] C. CHAUVE, J. COURTIEL, Y. PONTY. *Counting, generating and sampling tree alignments*, in "ALCOB - 3rd International Conference on Algorithms for Computational Biology - 2016", Trujillo, Spain, June 2016, https://hal.inria.fr/hal-01154030

[12] J. LUMBROSO, M. MISHNA, Y. PONTY. *Taming reluctant random walks in the positive quadrant*, in "GASCOM - 10th conference on random generation of combinatorial structures - 2016", Bastia, France, Electronic Notes in Discrete Mathematics, June 2016, forthcoming, https://hal.inria.fr/hal-01291164

National Conferences with Proceedings

[13] V. J. HENRY, A. FERRÉ, C. FROIDEVAUX, A. A. GOELZER, V. V. FROMION, S. COHEN-BOULAKIA, S. S. DEROZIER, M. DINH, G. FIÉVET, S. FISCHER, J.-F. J.-F. GIBRAT, V. L. LOUX, S. PÉRÈS. *Représentation*

systémique multi-échelle des processus biologiques de la bactérie, in "IC2016: Ingénierie des Connaissances", Montpellier, France, June 2016, https://hal.archives-ouvertes.fr/hal-01442727

[14] A. SAAIDI, D. ALLOUCHE, B. SARGUEIL, Y. PONTY. *Towards structural models for the Ebola UTR regions using experimental SHAPE probing data*, in "JOBIM - Journées Ouvertes en Biologie, Informatique et Mathématiques - 2016", Lyon, France, June 2016, https://hal.inria.fr/hal-01332469

[15] W. WANG, M. BARBA, P. RINAUDO, A. DENISE, Y. PONTY. *Homology modeling of complex structural RNAs*, in "JOBIM - Journées Ouvertes en Biologie, Informatique et Mathématiques - 2016", Lyon, France, June 2016, https://hal.inria.fr/hal-01332642

Other Publications

[16] C. BARTON, A. HELIOU, L. MOUCHARD, S. P. PISSIS. *Parallelising the Computation of Minimal Absent Words*, January 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01255489

[17] 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

[18] 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

[19] 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

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

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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] W. KLADWANG, R. DAS. *A mutate-and-map strategy for inferring base pairs in structured nucleic acids: proof of concept on a DNA/RNA helix*, in "Biochemistry", 2010, vol. 49, n^{o} 35, pp. 7414–7416

[31] 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

[32] 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

[33] 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

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

[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 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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]

[43] 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

[44] 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

[45] L. WEINBRAND, A. AVIHOO, D. BARASH. *RNAfbinv: an interactive Java application for fragment-based design of RNA sequences*, in "Bioinformatics", 2013, btt494