IN PARTNERSHIP WITH:
CNRS

Ecole Polytechnique

Université Paris-Sud (Paris 11)

Activity Report 2015

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

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
1.1.10. - Mathematical biology
1.1.9. - Bioinformatics
9.6. - Reproducibility

1. Members

Research Scientists

Mireille Regnier [Team leader, Inria, Senior Researcher, HdR]
Julie Bernauer [Inria, Researcher, HdR]
Yann Ponty [CNRS, Researcher]

Faculty Member

Philippe Chassignet [Ecole Polytechnique, Associate Professor]

PhD Students

Alice Heliou [Ecole Polytechnique]
Amélie Heliou [Ecole Polytechnique]
Daria Iakovishina [Ecole Polytechnique]
Vincent Le Gallic [Universite Paris XI, from May 2015]
Afaf Saaidi [Ecole Polytechnique]
Antoine Soule [Ecole Polytechnique]

Administrative Assistant

Evelyne Rayssac [Ecole Polytechnique]

Others

Loïc Paulevé [CNRS, Researcher]
Jean-Marc Steyaert [Ecole Polytechnique, émérite, HdR]

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

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, Antoine Soulé.

*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 [56], the correction of pyrosequencing errors [48], and the efficient detection of multi-stable RNAs (riboswitches) [50], [51].

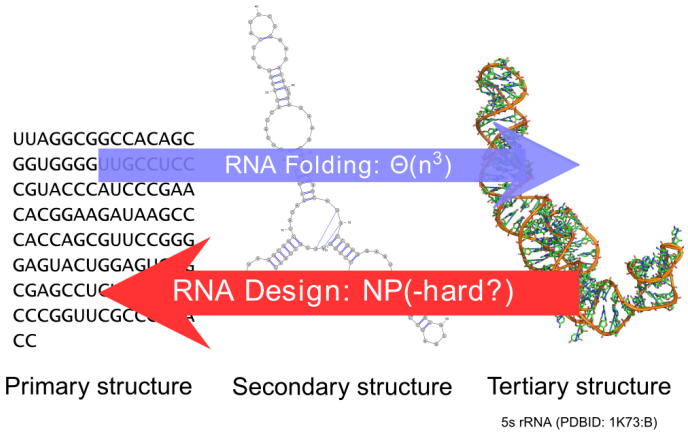

Figure 1. The goal of RNA design, aka RNA inverse folding, is to find a sequence that folds back into a given (secondary) structure.

3.1.2. RNA design.

Participants: Alice Heliou, Vincent Le Gallic, Yann Ponty.

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

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

Within our group, we offer a new methodology, based on weighted random generation [33] 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 [39].

3.1.3. Towards 3D modeling of large molecules

Participants: Yann Ponty, Afaf Saaidi, Mireille Regnier.

Joint projects with A. Denise (sc 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. Sequences

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

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

As a whole, motif detection in genomic sequences is a hot subject in computational biology that allows to address some key questions such as chromosome dynamics or annotation. 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 [44]. 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, Daria Iakovishina.

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

Motif detection combines an algorithmic search of potential sites and a significance assessment. Assessment significance requires a quantitative criteriumsuch 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 [47]. Recursive equations to compute pvalues 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 [54], 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 [37], and, recently, heuristics have been proposed and implemented [34], [49], [30]. 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 [29], [32]. 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 [52]. We have studied the computation of minimal absent words and we have provided new linear implementations [25],[21].

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 [27]. 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

Participant: Yann Ponty.

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

In 2005, a new paradigm appeared in the *ab initio* secondary structure prediction [35]: 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 [46], [5], including combinatorial models for structures with pseudoknots. Our aim is to combine this paradigm with a fragment based approach for decomposing structures, such as the cycle decomposition used within F. Major's group [44].

3.3. 3D interaction and structure prediction

Participants: Julie Bernauer, 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 [26], [36] and efficient code are essential [40].

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.

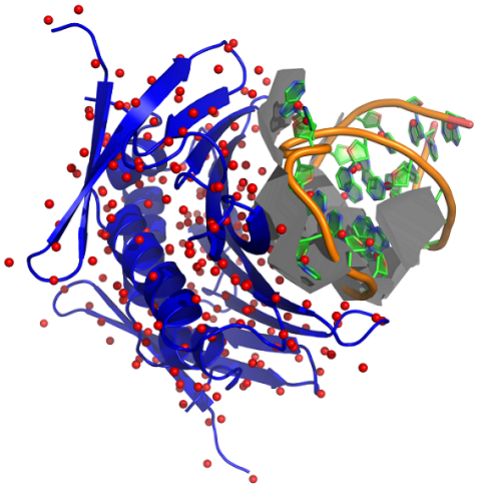

Figure 2. Coarse-grained representation and Voronoi interface model of a PP7 coat protein bound to an RNA hairpin (PDB code 2qux). The Voronoi model captures the features of the interactions such as stacking, even at the coarse-grained level.

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

Participants: Julie Bernauer, Amélie Heliou.

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

Our current and future efforts towards knowledge-based scoring function and ion location prediction have been described in 3.3.1.

Over the last two decades a strong connection between robotics and computational structural biology has emerged, in which internal coordinates of proteins are interpreted as a kinematic linkage with rotatable bonds as joints and corresponding groups of atoms as links [55], [31], [43], [42]. Initially, fragments in proteins limited to tens of residues were modeled as a kinematic linkage, but this approach has been extended to

encompass (multi-domain) proteins [41]. For RNA, progress in this direction has been realized as well. A kinematics-based conformational sampling algorithm, KGS, for loops was recently developed [38], but it does not fully utilize the potential of a kinematic model. It breaks and recloses loops using six torsional degrees of freedom, which results in a finite number of solutions. The discrete nature of the solution set in the conformational space makes difficult an optimization of a target function with a gradient descent method. Our methods overcome this limitation by performing a conformational sampling and optimization in a co-dimension 6 subspace. Fragments remain closed, but these methods are limited to proteins. Our objective is to extend the approach proposed in [38], [55] to nucleic acids and protein/nucleic acid complexes with a view towards improving structure determination of nucleic acids and their complexes and in silico docking experiments of protein/RNA complexes. For that purpose, we have developed a generic strategy for differentiable statistical potentials [1], [53] that can be directly integrated in the procedure.

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

4. Highlights of the Year

4.1. Highlights of the Year

4.1.1. Keynote addresses

Y. Ponty delivered one of the 8 plenary addresses at the 5th biennial Canadian Discrete and Algorithmic Mathematics Conference (CanaDAM) in University of Saskatchewan (Saskatoon, Canada). Held every two-years, with ∼300 participants and ∼150 contributed and invited talks, CanaDAM is the foremost event in Discrete Mathematics in Canada.

4.1.2. Awards

Alice Héliou received "Prix Poster École Doctorale Interfaces,Pôle : Science Du Vivant"

5. New Software and Platforms

5.1. VARNA

KEYWORDS: Bioinformatics - Structural Biology
FUNCTIONAL DESCRIPTION

A lightweight Java Applet dedicated to the quick drawing of an RNA secondary structure. VARNA is open-source and distributed under the terms of the GNU GPL license. Automatically scales up and down to make the most out of a limited space. Can draw multiple structures simultaneously. Accepts a wide range of documented and illustrated options, and offers editing interactions. Exports the final diagrams in various file formats (svg,eps,jpeg,png,xfig)

- Participants: Yann Ponty
- Contact: Yann Ponty
- URL: http://varna.lri.fr/

5.2. KGS

KEYWORDS: Bioinformatics - Structural Biology -protein kinematics -RNA kinematics

FUNCTIONAL DESCRIPTION

The Kino-Geometric Sampling (KGS) software suite uses advanced, robotics-inspired algorithms to rapidly explore the conformational landscape of folded proteins, RNA, and their complexes. Combined with powerful statistical techniques, it structurally characterizes collective motions and excited substates from sparse, spatiotemporally averaged data.

- Participants: Amélie Héliou.
- Contact: Amélie Héliou
- URL: https://simtk.org/home/kgs/

5.3. SV-BAY

KEYWORDS: Bioinformatics - NGS- Cancer
FUNCTIONAL DESCRIPTION

SV-BAY is a software to detect structural variants in cancer genomes. It relies on a Bayesian approach and a correction for GC-content and read mappability is provided. SV-BAY is written in Python with small insertions in C++ code.

- Participants: Daria Iakovishina and M. Régnier
- Contact: M. Régnier
- URL: https://github.com/InstitutCurie/SV-Bay

5.4. DeClone

KEYWORDS: Bioinformatics - Comparative Genomics - Genome rearrangements
FUNCTIONAL DESCRIPTION

DECLONE is a software to predict ancestral adjacencies from reconciled gene trees. It offers multiple indicators to assess the robustness of predictions, including individual supports, the (stochastic) generation of (co/sub)-optimal solutions, and the domain of validity of a given prediction in the parameter space.

- Participants: Y. Ponty
- Contact: Y. Ponty
- URL: https://github.com/yannponty/DeClone

6. New Results

6.1. RNA Design

In collaboration with J. Hales, J. Manuch and L. Stacho (Simon Fraser University/Univ. British Columbia, Canada), we have investigated the combinatorial RNA design problem, a minimal instance of the RNA design problem which aims at finding a sequence that admits a given target as its unique base pair maximizing structure. We obtained provide 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 provides 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.

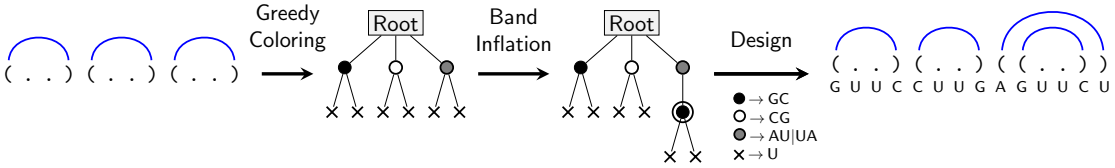

Figure 3. Principle of our structure-approximating version of RNA design: Starting from a potentially undesignable structure, a greedy coloring can be performed and corrected such that the final structure is provably designable in linear time.

Theses results were presented at the CPM 2015 conference in Italy [17], and open new avenues of research, both towards pratical, tractable versions of design, and constitute a first step towards long-awaited theoretical foundations for the problem.

6.2. Combinatorics of motifs and algorithms

We developed an $O(n)$-time and $O(n)$-space algorithm to compute minimal absent words. Their computation is used in sequence comparison [32] or to detect biologically significant events. For instance, in [52], it was shown that there exist three minimal words in *Ebola* virus genomes which are absent from human genome. The identification of such species-specific sequences may prove to be useful for the development of both diagnosis and therapeutics. In our new contribution [21] we provided an implementation that can be executed in parallel. Experimental resuts show that excluding the indexing data structure construction time, it achieves near-optimal speed-ups. The computation on the human genome is accelerated by a factor of 10 when using 16 processors, but it consummes a huge amout of RAM. Thus we are currently working on an external memory implementation, that will provide a trade-off between space and time consumption.

Combinatorial tools have been developed to predict the length of repetitions in a random sequence. This allows to distinguish biologically significant repetitions or tune some parameters in assembly or re-sequencing algorithms. For instance, unique mappability is strongly related to the length of the repetitions. A *trie profile* was defined in [45] to address this issue for binary alphabets, by the means of analytic combinatorics. General alphabets, where no closed formula exist, were adressed in [24]. An alternative, and simpler, approach is derived, thatexhibits a Large deviation Principle and makes use of Lagrange multipliers. Different domains and transition phases are exhibited. It is expected that htis approach extends to a Markov model and to approximate repetitions.

6.3. Structural variants

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 [12] 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.

Structural variations can also be observed and analyzed at larger time scales, and computational methods can be used to predict the structure of ancestral genomes. Within two collaborations with C. Chauve, A. Rajaraman (Simon Fraser University, Canada) and J. Zanetti (SFU, Canada & UniCAMP, Brazil), we revisited the problem of predicting a parsimonious set of adjacencies between ancestral genes, i.e. the most likely structure of an ancestral genome. More specifically, we modified the dynamic programming scheme underlying the DeCo algorithm [28] to compute indicators of robustness for predicting adjacencies. Our reimplementation, which relies on interesting meta-programming strategies, is available at https://github.com/yannponty/DeClone.

In a first study, we postulated a Boltzmann-Gibbs distribution over the set of evolutionary scenarii [9]. Our initial experiments relied on Boltzmann sampling to estimate the probabilities of ancestral adjacencies, but our extended version describes an exact polynomial-time computation of such probabilities, through an adaptation of the inside-outside algorithm. We interpreted such probabilities as supports for predicted adjacencies, and found that discarding adjacencies associated with low supports provided a good strategy for resolving synthenic conflicts.

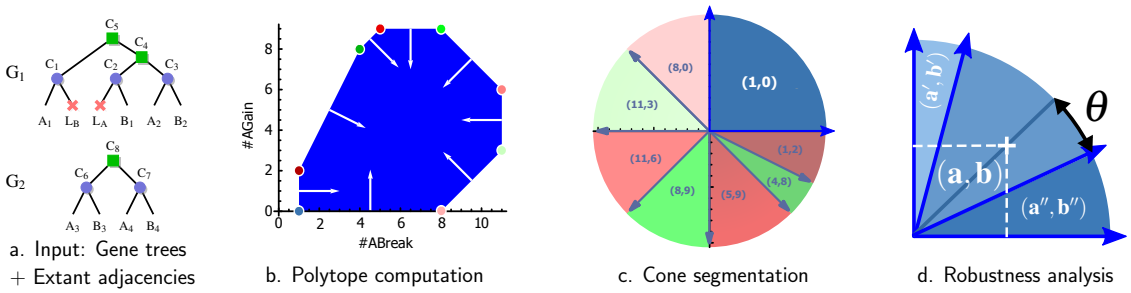

Figure 4. Main steps involved in the parametric prediction of ancestral adjacencies. Starting from two reconciled gene trees and a list of contemporary adjacencies (a.), the polytope of admissible Adjacency Gains/Breaks (+Presence/Absence of a given adjacency) is computed (b.) and projected onto a dual space which partitions the space of cost schemes into (infinite) regions leading to equivalent predictions (c.). The angular distance of the reference cost scheme $(1, 1)$ to a region representing an alternative prediction (d.) is used as a measure of robustness for the prediction.

However, the costs associated with the main operations (gaining/breaking adjacencies) in the underlying evolutionary models must be set beforehand in a somewhat arbitrary fashion. This has led us to investigate the influence of those costs on the characteristics of parsimonious predictions, i.e. the robustness of predictions with respect to perturbations of the scoring scheme [18]. To that purpose, we have performed an exact parametric analysis of the DeCo dynamic programming scheme (see Fig. 4 for details). This analysis revealed a quasi-independence, for a large subset of gene trees, of predicted adjacencies to the actual numerical values involved in the scoring scheme.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. FRM

Y. 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
- Y. Ponty, A. Denise, M. Regnier, A. Saaidi (PhD funded by FRM)
- B. Sargueil (Paris V – Experimental partner), J. Waldispuhl (Univ. McGill)

7.2. European Initiatives

Y. 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).

ANR International Program

- Fast and efficient sampling of structures in RNA folding landscapes
- RNALANDS (ANR-14-CE34-0011)
- 01/10/2014–30/09/2018
- Y. Ponty (PI), M. Régnier
- EPI BONSAI/INRIA Lille - Nord Europe, Vienna University (Austria)
- LRI, Université Paris-Sud (France)

7.3. International Initiatives

7.3.1. Inria International Partners

7.3.1.1. Declared Inria International Partners

AMAVI

Title: Combinatorics and Algorithms for the Genomic sequences

International Partners (Institution - Laboratory - Researcher):

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

Start year: 2013

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

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.

7.3.1.2. Informal International Partners

A long-term cooperation exists with Teheran University (Iran).

7.3.2. Participation In other International Programs

CONSEIL FRANCO-QUÉBÉCOIS DE COOPÉRATION UNIVERSITAIRE EXCHANGE PROGRAM

- Title: Réseau franco-québecois de recherche sur l'ARN
- International Partners (Institution - Laboratory - Researcher):
  Univ. McGill (Canada) - CS Dept - J. Waldispühl, M. Blanchette
  Univ. Montréal (Canada) - Biology Dept & IRIC - E. Lecuyer, F. Major
- Start year: 2012
- The partners have developped complementary expertise on RNA : bioinformatics, combinatorics and algorithms. machine learning, physics and genomics. Methodologies will be developed that combine theoretical simulations and new (high throughput) experimental data. A common high level training at Master and PhD level is organized.

PHC GERMAINE DE STAEL EXCHANGE PROGRAM

- Title: Random constrained permutations
- International Partners (Institution - Laboratory - Researcher):
  Univ. Zürich (Swiss) - Institut für Mathematik - M. Bouvel, V. Féray
- Start year: 2015
- The partners wish to develop new technique for the enumeration, analysis and random generation of constrained permutations.

CNRS UMI PIMS-VANCOUVER EXCHANGE PROGRAM

- Title: Extended research stay of Y. Ponty at the Simon Fraser University
  Simon Fraser University - Maths Dept - C. Chauve, M. Mishna, L. Stacho
  Univ. British Columbia - CS Dept - J. Manuch
- Start year: 2013
- Extended research stay in Vancouver to foster new collaborations between EPI Amib and colleagues at SFU on comparative genomics, RNA structures, and enumerative combinatorics.

7.4. International Research Visitors

7.4.1. Visits of International Scientists

Mark Ward
Date: 23/11/2015- 05/12/2015
Institution: Purdue University (USA)

Can Alkan
Date: 24/11/2015- 30/11/2015
Institution: Bilkent University (Turkey)

Evgenia Furletova
Date: 22/11/2015- 28/11/2015
Institution: IMPB (Russia)

7.4.1.1. Internships

Indrajit Saha
Date: 20/02/2015- 28/02/2015
Institution:ERCIM fellowship (Wroclaw)
Supervisor: M. Régnier

Doris Taining

Date: 01/05/2015- 07/08/2015

Institution:Singapore University (Singapore)

Supervisor: M. Régnier

7.4.2. Visits to International Teams

7.4.2.1. Sabbatical programme

Bernauer Julie

Date: Feb 2014 - Jan 2015

Institution: Stanford (United States)

7.4.2.2. Research stays abroad

Yann Ponty

Date: Sept 2013 - Sept 2015

Institution: Simon Fraser (Canada)

Amelie Héliou

Date:June 2015 -Aug 2015

Institution: HKUST (Hong Kong)

Antoine Soulé

Date: Jan 2015 - Sept 2015

Institution: McGill (Canada)

Pauline Pommeret

Date: May 2015 - Aug 2015

Institution: Vancouver (Canada)

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific events organisation

8.1.1.1. General chair, scientific chair

Mireille Régnier

SeqBio 2015

NGS day at INRIA

8.1.1.2. Member of the organizing committees

Yann Ponty

SeqBio 2015

8.1.2. Scientific events selection

8.1.2.1. Chair of conference program committees

Mireille Régnier acted as chair of the program committee for the SeqBio 2015 workshop (Orsay, France).

8.1.2.2. Member of the conference program committees

Yann Ponty

ISMB/ECCB 2015

WABI 2015

BioVis 2015

BiCOB 2015

SeqBio 2015

Mireille Régnier

SeqBio 2015

MCCMB 2015

8.1.2.3. Reviewer

Y. Ponty acted as an external reviewer for the ACM-SIGMOD 2015 conference.

8.1.3. Journal

8.1.3.1. Member of the editorial boards

M. Régnier is an editor of PeerJ Computer Science.

8.1.3.2. 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...

8.1.4. Invited talks

Y. Ponty:

Keynote address at the bi-annual the Canadian Discrete Applied Mathematics (CanaDAM'15) meeting held at the university of Saskatoon (Canada);

Invited talk at the *Benasque RNA* triannual meeting (Spain);

Invited talk at the Mathematical Biosciences Institute (Columbus, Ohio) during the *Geometric and Topological Modeling of Biomolecules workshop*;

Invited talk at the bi-annual *meeting of the CNRS GDR BIM* at Institut Pasteur, Paris.

M. Régnier

Invited talk at Sydney University (Australia).

8.1.5. Leadership within the scientific community

Y. Ponty is scientific animator (2014-2018) of the *Structure and interactions of macromolecules* axis of the CNRS research group in molecular bioinformatics (GdR BIM);

Y. Ponty and M. Régnier are active members of the COMATEGE and the ALEA working group http://igm.univ-mlv.fr/~josuatv/webalea/ of the CNRS research group in Mathematical Computer Science (GDR IM) https://www.gdr-im.fr/.

8.1.6. Scientific expertise

Y. Ponty is a member of the 'Comité National' (hiring/evaluation committee) of CNRS in computer science (section 6) and Maths/Physics/Computer Science interfaces with life science (CID 51); he acted as an external expert for the Emergence program of Ville de Paris

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

8.1.7. Research administration

M. Regnier participates to the *Conseil de laboratoire* of LIX as head of the AMIB team.

Y. Ponty is an elected member of the *comité national du CNRS*, and takes 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).

8.2. Teaching - Supervision - Juries

8.2.1. 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. We are also involved in a student partnership with McGill University (partenariat France Quebec offering French and Canadian students co-supervised internships (short term -3 to 6 months- or long term -part of the PhD studies-).

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

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

8.2.2. Supervision

HdR

Julie Bernauer, *Geometric and statistical methods for the analysis and prediction of structural interactions between biomolecules*, Université Paris-Sud XI, January 2015, Habilitation à diriger des recherches. https://tel.archives-ouvertes.fr/tel-01136261

PhD

Daria Iakovishina, *Detection of structural variants in tumoral genomes with a Bayesian approach*, Ecole Polytechnique, November 2015. Encadrantes: Mireille Régnier and Valentina Boeva.

PhD in progress

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

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

Amélie Heliou, *Game theory and conformation sampling for multi-scale and multi-body macromolecule docking*, Ecole Polytechnique, Encadrant(els): Johanne Cohen

Vincent Le Gallic, *Design de structures secondaires avec contraintes de séquences : une approche globale fondée sur les langages formels*, Univ. Paris XI, Encadrant(els): Yann Ponty and Alain Denise.

Cécile Pereira, *Nouvelles approches bioinformatiques pour l'étude à grande échelle de l'évolution des activités enzymatiques*, Univ. Paris XI, Encadrant(els): Olivier Lespinet et Alain Denise.

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)

8.2.3. Juries

Yann, JMS, Philippe, MR

Hiring committees

Chargé de Recherches,INRIA, 2015, CRI Saclay Ile de France: M. Régnier

Ingénieur,INRIA, 2015, CRI Grenoble Rhone-Alpes: Mireille Régnier

Research Scientists/Directors, CNRS, Theoretical Computer Science (section 6) and Interfaces of Life Sciences (CID 51), 24 positions in 2015: Y. Ponty

PhD juries

Paul Dallaire, RNA Bioinformatics, Université de Montréal: Y. Ponty (External reviewer)

Magali Semeria, Comparative Genomics, Université Lyon I: Y. Ponty (Jury Member)

HDR juries

J. Bernauer (Paris-Sud U.) : J.-M. Steyaert

8.3. Popularization

AMIB animated the *Construisons les ARN* booth at *Fête de la Science*, October 2015, http://www.inria.fr/centre/saclay/agenda/fete-de-la-science-2015 with the participation of Y. Ponty, M. Régnier, A. Héliou, A. Héliou and A. Saaidi.

9. Bibliography

Major publications by the team in recent years

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

[2] J. HALEŠ, J. MAŇUCH, Y. PONTY, L. STACHO. *Combinatorial RNA Design: Designability and Structure-Approximating Algorithm*, in "CPM - 26th Annual Symposium on Combinatorial Pattern Matching", Ischia Island, Italy, F. CICALESE, E. PORAT (editors), June 2015, Accepted (To appear), https://hal.inria.fr/hal-01115349

[3] 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, pp. i308-15, Extended version of ISMB/ECCB'13 [*DOI :* 10.1093/BIOINFORMATICS/BTT217], https://hal.inria.fr/hal-00840260

[4] M. RÉGNIER, E. FURLETOVA, V. YAKOVLEV, M. ROYTBERG. *Analysis of pattern overlaps and exact computation of P-values of pattern occurrences numbers: case of Hidden Markov Models*, in "Algorithms for Molecular Biology", December 2014, vol. 9, n^{o} 1 [*DOI :* 10.1186/s13015-014-0025-1], https://hal.inria.fr/hal-00858701

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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[6] J. BERNAUER. *Geometric and statistical methods for the analysis and prediction of structural interactions between biomolecules*, Université Paris-Sud XI, January 2015, Habilitation à diriger des recherches, https://tel.archives-ouvertes.fr/tel-01136261

Articles in International Peer-Reviewed Journals

[7] M. BOUDARD, J. BERNAUER, D. BARTH, J. COHEN, A. DENISE. *GARN: Sampling RNA 3D Structure Space with Game Theory and Knowledge-Based Scoring Strategies*, in "PLoS ONE", August 2015, vol. 10, n° 8, e0136444, https://hal.archives-ouvertes.fr/hal-01201665

[8] T. BOURQUARD, F. LANDOMIEL, E. REITER, P. CRÉPIEUX, D. W. RITCHIE, J. AZÉ, A. POUPON. *Unraveling the molecular architecture of a G protein-coupled receptor/β-arrestin/Erk module complex*, in "Scientific Reports", June 2015, 5:10760 [*DOI :* 10.1038/SREP10760], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01162594

[9] C. CHAUVE, Y. PONTY, J. P. P. ZANETTI. *Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach*, in "BMC Bioinformatics", December 2015, vol. 16, n° Suppl 19, S6 p. [*DOI :* 10.1186/1471-2105-16-S19-S6], https://hal.inria.fr/hal-01245495

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

[11] M. HEINONEN, O. GUIPAUD, F. MILLIAT, V. BUARD, B. MICHEAU, G. TARLET, M. BENDERITTER, F. ZEHRAOUI, F. D'ALCHÉ-BUC. *Detecting time periods of differential gene expression using Gaussian processes: An application to endothelial cells exposed to radiotherapy dose fraction*, in "Bioinformatics", March 2015, vol. 31, n° 5, pp. 728–735 [*DOI :* 10.1093/BIOINFORMATICS/BTU699], https://hal.archives-ouvertes.fr/hal-01154010

[12] 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

[13] H. JIANG, F. K. SHEONG, L. ZHU, X. GAO, J. BERNAUER, X. HUANG. *Markov State Models Reveal a Two-Step Mechanism of miRNA Loading into the Human Argonaute Protein: Selective Binding followed by Structural Re-arrangement*, in "PLoS Computational Biology", June 2015, vol. 11, n° 7, e1004404, https://hal.inria.fr/hal-01257437

[14] J. D. V. MOREIRA, S. PÉRÈS, J.-M. STEYAERT, E. BIGAN, L. PAULEVÉ, M. L. NOGEIRA, L. SCHWARTZ. *Cell cycle progression is regulated by intertwined redox oscillators*, in "Theoretical Biology and Medical Modelling", May 2015, vol. 12, n° 1, 10 p. [*DOI :* 10.1186/S12976-015-0005-2], https://hal.archives-ouvertes.fr/hal-01158514

International Conferences with Proceedings

[15] J. BERNAUER, R. FONSECA, H. VAN DEN BEDEM. *KGSrna: Efficient 3D Kinematics-Based Sampling for Nucleic Acids*, in "RECOMB 2015", Warsaw, Poland, Research in Computational Molecular Biology, April 2015, vol. 9029 [*DOI :* 10.1007/978-3-319-16706-0_11], https://hal.inria.fr/hal-01257755

[16] B. BRANCOTTE, B. YANG, G. BLIN, S. COHEN-BOULAKIA, A. DENISE, S. HAMEL. *Rank aggregation with ties: Experiments and Analysis*, in "The 41st International Conference on Very Large Data Bases", Kohala Coast, Hawaï, United States, August 2015, vol. 8, n° 11, https://hal.archives-ouvertes.fr/hal-01152098

[17] J. HALEŠ, J. MAŇUCH, Y. PONTY, L. STACHO. *Combinatorial RNA Design: Designability and Structure-Approximating Algorithm*, in "CPM - 26th Annual Symposium on Combinatorial Pattern Matching", Ischia Island, Italy, F. CICALESE, E. PORAT (editors), June 2015, forthcoming, https://hal.inria.fr/hal-01115349

[18] A. RAJARAMAN, C. CHAUVE, Y. PONTY. *Assessing the robustness of parsimonious predictions for gene neighborhoods from reconciled phylogenies*, in "ISBRA - 11th International Symposium on Bioinformatics Research and Applications - 2015", Norfolk, Virginia, United States, June 2015, forthcoming, https://hal.inria.fr/hal-01104587

Scientific Books (or Scientific Book chapters)

[19] F. JOSSINET, Y. PONTY, J. WALDISPÜHL. *Preface*, in "Computational methods for Structural RNAs (CMSR'14)", Journal of Computational Biology, Mary Ann Liebert, February 2015, vol. 22, n^{o} 3, 189 p. , https://hal.inria.fr/hal-01136745

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

Other Publications

[21] 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

[22] C. CHAUVE, J. COURTIEL, Y. PONTY. *Counting, generating and sampling tree alignments*, 2015, Submitted, https://hal.inria.fr/hal-01154030

[23] A. HELIOU, M. LÉONARD, L. MOUCHARD, M. SALSON. *Efficient Dynamic Range Minimum Query*, January 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01255499

[24] M. REGNIER, P. CHASSIGNET. *Accurate prediction of the statistics of repetitions in random sequences : a case study in Archae genomes*, 2015, working paper or preprint, https://hal.inria.fr/hal-01253628

References in notes

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

[26] 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

[27] 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

[28] S. BÉRARD, C. GALLIEN, B. BOUSSAU, G. J. SZÖLLÖSI, V. DAUBIN, E. TANNIER. *Evolution of gene neighborhoods within reconciled phylogenies*, in "Bioinformatics", 2012, vol. 28, n° 18, pp. i382-i388 [*DOI :* 10.1093/BIOINFORMATICS/BTS374], http://bioinformatics.oxfordjournals.org/content/28/18/i382.abstract

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

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

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

[32] 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

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

[34] H. DEVILLERS, S. SCHBATH. *Separating significant matches from spurious matches in DNA sequences*, in "Journal of Computational Biology", 2012, vol. 19, n° 1, pp. 1–12 [*DOI :* 10.1089/CMB.2011.0070]

[35] 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

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

[37] 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° 1-2, pp. 81–88

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

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

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

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

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

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

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

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

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

[47] 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

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

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

[50] 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° 12 [*DOI :* 10.1371/JOURNAL.PONE.0050506], http://hal.inria.fr/hal-00769740

[51] 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

[52] 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]

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

[54] 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

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

[56] 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