IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2017

Project-Team BONSAI

Bioinformatics and Sequence Analysis

IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille

RESEARCH CENTER
Lille - Nord Europe

THEME
Computational Biology

Table of contents

Project-Team BONSAI

Creation of the Project-Team: 2011 January 01

Keywords:

Computer Science and Digital Science:

A7.1. - Algorithms
A8.1. - Discrete mathematics, combinatorics
A8.7. - Graph theory

Other Research Topics and Application Domains:

B1.1.6. - Genomics
B1.1.7. - Immunology
B1.1.8. - Evolutionnary biology
B1.1.9. - Bioinformatics
B1.1.13. - Plant Biology
B1.1.14. - Microbiology
B2.2.3. - Cancer

1. Personnel

Research Scientists

Hélène Touzet [Team leader, CNRS, Senior Researcher, HDR]
Samuel Blanquart [Inria, Researcher]
Rayan Chikhi [CNRS, Researcher]
Mathieu Giraud [CNRS, Researcher, HDR]

Faculty Members

Stéphane Janot [Univ des sciences et technologies de Lille, Associate Professor]
Laurent Noé [Univ des sciences et technologies de Lille, Associate Professor]
Maude Pupin [Univ des sciences et technologies de Lille, Associate Professor, HDR]
Mikaël Salson [Univ des sciences et technologies de Lille, Associate Professor]
Jean-Stéphane Varré [Univ des sciences et technologies de Lille, Professor, HDR]

Post-Doctoral Fellows

Aymeric Antoine-Lorquin [Inria, from Oct 2017]
Benjamin Momège [Inria, until Feb 2017]

PhD Students

Quentin Bonenfant [CNRS, from Nov 2017]
Yoann Dufresne [Univ des sciences et technologies de Lille, until Feb 2017]
Pierre Marijon [Inria]
Pierre Pericard [Univ des sciences et technologies de Lille, until Oct 2017]
Tatiana Rocher [Univ des sciences et technologies de Lille]
Chadi Saad [CHRU Lille]
Léa Siegwald [until Feb 2017]

Technical staff

Aurélien Béliard [CHRU Lille, until Nov 2017]
Areski Flissi [CNRS]
Ryan Herbert [Inria]

Maël Kerbiriou [Inria, from Oct 2017]

Administrative Assistant

Amélie Supervielle [Inria]

2. Overall Objectives

2.1. Presentation

BONSAI is an interdisciplinary project whose scientific core is the design of efficient algorithms for the analysis of biological macromolecules.

From a historical perspective, research in bioinformatics started with string algorithms designed for the comparison of sequences. Bioinformatics became then more diversified and by analogy to the living cell itself, it is now composed of a variety of dynamically interacting components forming a large network of knowledge: Systems biology, proteomics, text mining, phylogeny, structural biology, etc. Sequence analysis still remains a central node in this interconnected network, and it is the heart of the BONSAI team.

It is a common knowledge nowadays that the amount of sequence data available in public databanks grows at an exponential pace. Conventional DNA sequencing technologies developed in the 70's already permitted the completion of hundreds of genome projects that range from bacteria to complex vertebrates. This phenomenon is dramatically amplified by the recent advent of Next Generation Sequencing (NGS), that gives rise to many new challenging problems in computational biology due to the size and the nature of raw data produced. The completion of sequencing projects in the past few years also teaches us that the functioning of the genome is more complex than expected. Originally, genome annotation was mostly driven by protein-coding gene prediction. It is now widely recognized that non-coding DNA plays a major role in many regulatory processes. At a higher level, genome organization is also a source of complexity and have a high impact on the course of evolution.

All these biological phenomena together with big volumes of new sequence data provide a number of new challenges to bioinformatics, both on modeling the underlying biological mechanisms and on efficiently treating the data. This is what we want to achieve in BONSAI. For that, we have in mind to develop well-founded models and efficient algorithms. Biological macromolecules are naturally modeled by various types of discrete structures: String, trees, and graphs. String algorithms is an established research subject of the team. We have been working on spaced seed techniques for several years. Members of the team also have a strong expertise in text indexing and compressed index data structures, such as BWT. Such methods are widely-used for the analysis of biological sequences because they allow a data set to be stored and queried efficiently. Ordered trees and graphs naturally arise when dealing with structures of molecules, such as RNAs or non-ribosomal peptides. The underlying questions are: How to compare molecules at structural level, how to search for structural patterns ? String, trees and graphs are also useful to study genomic rearrangements: Neighborhoods of genes can be modeled by oriented graphs, genomes as permutations, strings or trees.

A last point worth mentioning concerns the dissemination of our work to the biology and health scientific community. Since our research is driven by biological questions, most of our projects are carried out in collaboration with biologists. A special attention is given to the development of robust software, its validation on biological data and its availability from the software platform of the team: http://bioinfo.lille.inria.fr/.

3. Research Program

3.1. Sequence processing for Next Generation Sequencing

As said in the introduction of this document, biological sequence analysis is a foundation subject for the team. In the last years, sequencing techniques have experienced remarkable advances with Next Generation Sequencing (NGS), that allow for fast and low-cost acquisition of huge amounts of sequence data, and outperforms

conventional sequencing methods. These technologies can apply to genomics, with DNA sequencing, as well as to transcriptomics, with RNA sequencing. They promise to address a broad range of applications including: Comparative genomics, individual genomics, high-throughput SNP detection, identifying small RNAs, identifying mutant genes in disease pathways, profiling transcriptomes for organisms where little information is available, researching lowly expressed genes, studying the biodiversity in metagenomics. From a computational point of view, NGS gives rise to new problems and gives new insight on old problems by revisiting them: Accurate and efficient remapping, pre-assembling, fast and accurate search of non exact but quality labeled reads, functional annotation of reads, ...

3.2. Noncoding RNA

Our expertise in sequence analysis also applies to noncoding RNA. Noncoding RNA plays a key role in many cellular processes. First examples were given by microRNAs (miRNAs) that were initially found to regulate development in *C. elegans*, or small nucleolar RNAs (snoRNAs) that guide chemical modifications of other RNAs in mammals. Hundreds of miRNAs are estimated to be present in the human genome, and computational analysis suggests that more than 20% of human genes are regulated by miRNAs. To go further in this direction, the 2007 ENCODE Pilot Project provides convincing evidence that the Human genome is pervasively transcribed, and that a large part of this transcriptional output does not appear to encode proteins. All those observations open a universe of "RNA dark matter" that must be explored. From a combinatorial point of view, noncoding RNAs are complex objects. They are single stranded nucleic acid sequences that can fold forming long-range base pairings. This implies that RNA structures are usually modeled by complex combinatorial objects, such as ordered labeled trees, graphs or arc-annotated sequences.

3.3. Genome structures

Our third application domain is concerned with the structural organization of genomes. Genome rearrangements are able to change genome architecture by modifying the order of genes or genomic fragments. The first studies were based on linkage maps and fifteen year old mathematical models. But the usage of computational tools was still limited due to the lack of data. The increasing availability of complete and partial genomes now offers an unprecedented opportunity to analyze genome rearrangements in a systematic way and gives rise to a wide spectrum of problems: Taking into account several kinds of evolutionary events, looking for evolutionary paths conserving common structure of genomes, dealing with duplicated content, being able to analyze large sets of genomes even at the intraspecific level, computing ancestral genomes and paths transforming these genomes into several descendant genomes.

3.4. Nonribosomal peptides

Lastly, the team has been developing for several years a tight collaboration with ProBioGEM team in Institut Charles Viollette on nonribosomal peptides, and has became a leader on that topic. Nonribosomal peptide synthesis produces small peptides not going through the central dogma. As the name suggests, this synthesis uses neither messenger RNA nor ribosome but huge enzymatic complexes called Nonribosomal peptide synthetases (NRPSs). This alternative pathway is found typically in bacteria and fungi. It has been described for the first time in the 70's. For the last decade, the interest in nonribosomal peptides and their synthetases has considerably increased, as witnessed by the growing number of publications in this field. These peptides are or can be used in many biotechnological and pharmaceutical applications (e.g. anti-tumors, antibiotics, immuno-modulators).

4. Application Domains

4.1. Life Sciences and health

Our research plays a pivotal role in all fields of life sciences and health where genomic data are involved. This includes more specifically the following topics: plant genomics (genome structure, evolution, microRNAs), cancer (leukemia, mosaic tumors), drug design (NRPSs), environment (metagenomics and metatranscriptomics), virology (evolution, RNA structures) ...

5. Highlights of the Year

5.1. Highlights of the Year

- Bonsai and close partners organized the French conference in bioinformatics, JOBIM, in Lille. More than 350 people attended to the conference.
- In the two last years, more than 2,000 samples of patients suffering leukaemia were analyzed with the Vidjil software developed in Bonsai with our partners. The VidjilNet consortium will be launched on January 1st 2018 within the Inria Foundation.

5.1.1. Awards

Pierre Pericard received the Best Oral Presentation Award from the SFBI for its talk on MATAM at the French bioinformatics conference JOBIM.

BEST PAPER AWARD:

[29]
P. PERICARD, Y. DUFRESNE, S. BLANQUART, H. TOUZET. *Reconstruction of full-length 16S rRNA sequences for taxonomic assignment inmetagenomics*, in "JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques", Lille, France, July 2017, https://hal.inria.fr/hal-01574629

6. New Software and Platforms

6.1. BCALM 2

KEYWORDS: Bioinformatics - NGS - Genomics - Metagenomics - De Bruijn graphs
SCIENTIFIC DESCRIPTION: BCALM 2 is a bioinformatics tool for constructing the compacted de Bruijn graph from sequencing data. It is a parallel algorithm that distributes the input based on a minimizer hashing technique, allowing for good balance of memory usage throughout its execution. It is able to compact very large datasets, such as spruce or pine genome raw reads in less than 2 days and 40 GB of memory on a single machine.
FUNCTIONAL DESCRIPTION: BCALM 2 is an open-source tool for dealing with DNA sequencing data. It constructs a compacted representation of the de Bruijn graph. Such a graph is useful for many types of analyses, i.e. de novo assembly, de novo variant detection, transcriptomics, etc. The software is written in C++ and makes extensive use of the GATB library.

- Participants: Antoine Limasset, Paul Medvedev and Rayan Chikhi
- Contact: Rayan Chikhi
- Publication: Compacting de Bruijn graphs from sequencing data quickly and in low memory
- URL: https://github.com/GATB/bcalm

6.2. NORINE

Nonribosomal peptides resource
KEYWORDS: Drug development - Knowledge database - Chemistry - Graph algorithmics - Genomics - Biology - Biotechnology - Bioinformatics - Computational biology

SCIENTIFIC DESCRIPTION: Since its creation in 2006, Norine remains the unique knowledgebase dedicated to non-ribosomal peptides (NRPs). These secondary metabolites, produced by bacteria and fungi, harbor diverse interesting biological activities (such as antibiotic, antitumor, siderophore or surfactant) directly related to the diversity of their structures. The Norine team goal is to collect the NRPs and provide tools to analyze them efficiently. We have developed a user-friendly interface and dedicated tools to provide a complete bioinformatics platform. The knowledgebase gathers abundant and valuable annotations on more than 1100 NRPs. To increase the quantity of described NRPs and improve the quality of associated annotations, we are now opening Norine to crowdsourcing. We believe that contributors from the scientific community are the best experts to annotate the NRPs they work on. We have developed MyNorine to facilitate the submission of new NRPs or modifications of stored ones.
FUNCTIONAL DESCRIPTION: Norine is a public computational resource with a web interface and REST access to a knowledge-base of nonribosomal peptides. It also contains dedicated tools : 2D graph viewer and editor, comparison of NRPs, MyNorine, a tool allowing anybody to easily submit new nonribosomal peptides, Smiles2monomers (s2m), a tool that deciphers the monomeric structure of polymers from their chemical structure.

- Participants: Areski Flissi, Juraj Michalik, Laurent Noé, Maude Pupin, Stéphane Janot, Valerie Leclère and Yoann Dufresne
- Partners: CNRS - Université Lille 1 - Institut Charles Viollette
- Contact: Maude Pupin
- Publications: Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing - Smiles2Monomers: a link between chemical and biological structures for polymers - Norine: a powerful resource for novel nonribosomal peptide discovery - NORINE: a database of nonribosomal peptides. - Bioinformatics Tools for the Discovery of New Nonribosomal Peptides
- URL: http://bioinfo.lille.inria.fr/NRP

6.3. Vidjil

High-Throughput Analysis of V(D)J Immune Repertoire
KEYWORDS: Cancer - Indexation - NGS - Bioinformatics - Drug development
SCIENTIFIC DESCRIPTION: Vidjil is made of three components: an algorithm, a visualization browser and a server that allow an analysis of lymphocyte populations containing V(D)J recombinations.

Vidjil high-throughput algorithm extracts V(D)J junctions and gathers them into clones. This analysis is based on a spaced seed heuristics and is fast and scalable, as, in the first phase, no alignment is performed with database germline sequences. Each sequence is put in a cluster depending on its V(D)J junction. Then a representative sequence of each cluster is computed in time linear in the size of the cluster. Finally, we perform a full alignment using dynamic programming of that representative sequence against the germline sequences.

Vidjil also contains a dynamic browser (with D3JS) for visualization and analysis of clones and their tracking along the time in a MRD setup or in an immunological study.
FUNCTIONAL DESCRIPTION: Vidjil is an open-source platform for the analysis of high-throughput sequencing data from lymphocytes. V(D)J recombinations in lymphocytes are essential for immunological diversity. They are also useful markers of pathologies, and in leukemia, are used to quantify the minimal residual disease during patient follow-up. High-throughput sequencing (NGS/HTS) now enables the deep sequencing of a lymphoid population with dedicated Rep-Seq methods and software.

- Participants: Florian Thonier, Marc Duez, Mathieu Giraud, Mikaël Salson, Ryan Herbert and Tatiana Rocher
- Partners: CNRS - Inria - Université de Lille - CHRU Lille
- Contact: Mathieu Giraud

- Publications: High-Throughput Immunogenetics for Clinical and Research Applications in Immunohematology: Potential and Challenges. - High-throughput sequencing in acute lymphoblastic leukemia: Follow-up of minimal residual disease and emergence of new clones - Diagnostic et suivi des leucémies aiguës lymphoblastiques (LAL) par séquençage haut-débit (HTS) - Multiclonal Diagnosis and MRD Follow-up in ALL with HTS Coupled with a Bioinformatic Analysis - A dataset of sequences with manually curated V(D)J designations - Vidjil: A Web Platform for Analysis of High-Throughput Repertoire Sequencing - Multi-loci diagnosis of acute lymphoblastic leukaemia with high-throughput sequencing and bioinformatics analysis - Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing - The predictive strength of next-generation sequencing MRD detection for relapse compared with current methods in childhood ALL.
- URL: http://www.vidjil.org

6.4. MATAM

Mapping-Assisted Targeted-Assembly for Metagenomics
KEYWORDS: Metagenomics - Genome assembling - Graph algorithmics
SCIENTIFIC DESCRIPTION: MATAM relies on the construction of a read overlap graph. Overlaps are computed using SortMeRNA. The overlap graph is simplified into relevant components related to specific and conserved regions. Components are assembled into contigs using SGA and contigs are finally assembled into scaffolds. The process yields nearly full length marker sequences with a very low error rate compared to the state of the art approaches. Taxonomic assignation of the obtained scaffolds is performed using the RDP classifier and is represented using Krona.
FUNCTIONAL DESCRIPTION: MATAM provides targeted genes assembly from the short metagenomic reads issued from environmental samples sequencing. Its default application focuses on the gold standard for species identification, 16S / 18S ribosomal RNA SSU genes. The produced gene scaffolds are highly accurate and suitable for precise taxonomic assignation. The software also provides a RDP classification for the reconstructed scaffolds as well as an estimation of the relative population sizes.

- Participants: Hélène Touzet, Pierre Pericard, Yoann Dufresne, Samuel Blanquart and Loïc Couderc
- Contact: Hélène Touzet
- Publication: MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes
- URL: https://github.com/bonsai-team/matam

7. New Results

7.1. Metagenomics

Reconstruction of phylogenetic marker genes. Accurate identification of organisms present within a community is essential to understanding the structure of an ecosystem. However, current HTS technologies generate short reads, such as Illumina reads, which makes it a difficult task. One possibility is to focus on assembly of taxonomic markers of interest, such as 16S ribosomal RNA. The PhD thesis of P. Pericard proposed an algorithm that is specifically dedicated to this problem. The method implements a stepwise process based on construction and analysis of a read overlap graph, which is built using read alignments (produced by SortMeRNA) and is decomposed into relevant connected components extracted from a compressed representation of the graph. It is able to recover full length 16S sequences with high precision assemblies ($\leq 0.1\%$ error rate). This work is published in the reference journal in the field [23] and the resulting software, MATAM, was released this spring. It is currently being tested in several labs [1]. This work received the Best Oral Presentation Award from the SFBI [2] this year [29].

[1] Tests of MATAM at MEDIS (INRA-Université Clermont Auvergne) for gene capture, Labgem (Genoscope) where it is on tracks to be integrated into the PathoTRACK-MicroScope platform dedicated to the human intestinal microbiome, and the Australian Centre for Ancient DNA (University of Adelaïde) for oral microbiome research.
[2] SFBI: Société Française de Bioinformatique

Metagenomics assembly. Another important task that could help taxonomic assignment is to reconstruct uncultured microbial strains and species for which the genome sequence is fully unknown. To this end, metagenomics mainly borrows techniques from classical genomics, i.e. from *de novo* assembly of isolate genomes. We built upon continuous methodological advances with our genomic assembler Minia, adding new data structures such as the minimal perfect hash function [26] and the compressed graph representation. We participated in 2015 in the CAMI metagenomic reconstruction challenge [3]. This challenged gathered a total of 17 international groups, and Minia performed among the top assembly methods. This result is reported in an article to appear in Nature Methods from the CAMI consortium [25]. We further presented a poster at RECOMB 2017.

Targeted metagenomics. Within the PhD thesis of L. Siegwald, we have participated to the design of a comprehensive evaluation protocol to compare computational pipelines to analyze 16S amplicons, and have studied the impact of different variables on the biological interpretation of results. This study included the following tools: CLARK, Kraken, Mothur, Qiime and One Codex. It has been the subject of an invited keynote at the international workshop Recent Computational Advances in Metagenomics (RCAM 2017) [4].

7.2. Nonribosomal peptides

We further investigate the NRPs produced by *Burkholderia*, focusing on the identification of new compounds implicated in biocontrol and pharmaceutical [19].

New functionalities have been added to Norine to query the SMILES field either by the query form and the REST service. We also continue our curating of Norine data by improving peptide annotations and validation submissions of new peptides.

7.3. High-throughput V(D)J repertoire analysis

Researches on high-throughput V(D)J repertoire analysis started in the group in 2012. We have developed Vidjil, a web platform dedicated to the analysis of lymphocyte populations. Starting from DNA sequences, uploaded by the user, Vidjil identifies and quantifies lymphocyte populations and provides an interactive visualization.

Seven European hospitals are now using Vidjil for their daily clinical practice. This year we published our experience of the minimal residual disease follow-up for acute lymphoblastic leukemia using Vidjil [24]. This is a first step towards using high-throughput sequencing and Vidjil for all the follow-up of the patients. We also participated to a joint publication with the EuroClonality-NGS consortium (see below).

Finally, we are working on transferring activities on platform development and user support. After meetings with several partners, we selected the Inria Foundation. The VidjilNet consortium (http://www.vidjil.net) will be launched in January 2018 within the InriaSoft action of the Foundation and will hire two engineers. VidjilNet will first gather hematology labs of French hospitals working on diagnosis and follow-up of acute lymphoblastic leukemia, and will be then extended to labs working on other pathologies as well as foreign labs.

7.4. RNA-Seq software benchmarking

Plenty of methods have been devised to analyze RNA-Seq data. Due to this large choice, it is a difficult task to determine what software is the best suited for a given question. To help in solving this problem, with colleagues at IRMB and in the SeqOne start-up in Montpellier, we devised a flexible benchmarking pipeline [18].

This pipeline is intended to be flexible enough to deal either with simulated or real data and to evaluate software on many possible aspects (mapping, splice detection, fusion detection, variant calling, and also on the post-analysis aspects such as gene quantification).

[3] CAMI challenge: https://data.cami-challenge.org/
[4] RCAM 2017: http://maiage.jouy.inra.fr/?q=fr/rcam2017

7.5. RNA folding landscape

Kinetics is key to understand many phenomena involving RNAs, such as co-transcriptional folding and riboswitches. Exact out-of-equilibrium studies induce extreme computational demands, leading state-of-the-art methods to rely on approximated kinetics landscapes, obtained using sampling strategies that strive to generate the key landmarks of the landscape topology. However, such methods are impeded by a large level of redundancy within sampled sets. Such a redundancy is uninformative, and obfuscates important intermediate states, leading to an incomplete vision of RNA dynamics.

Within the context of ANR RNAlands, we introduced RNANR, a new set of algorithms for the exploration of RNA kinetics landscapes at the secondary structure level. RNANR considers locally optimal structures, a reduced set of RNA conformations, in order to focus its sampling on basins in the kinetic landscape. Along with an exhaustive enumeration, RNANR implements a novel non-redundant stochastic sampling, and offers a rich array of structural parameters. Our tests on both real and random RNAs reveal that RNANR allows to generate more unique structures in a given time than its competitors, and allows a deeper exploration of kinetics landscapes [27].

7.6. Large-scale sequencing data indexing

Petabytes of DNA and RNA sequencing data are currently stored in online databases. It is currently possible to access these databases in two ways: 1) metadata queries, such as organism, instrument type, etc, and 2) download raw data. Due to the sheer size of the data, the web servers do not offer the possibility to search for sequences inside datasets. Such an operation would be invaluable to biology investigators, for example to determine which experiments contain an organism of interest, high expression of a certain transcript, a certain mutation, etc. Prior work exists for indexing sequencing data (Bloom Filter Tries, Sequence Bloom Trees), yet the performance remains prohibitive (either high memory usage, or several days for performing certain queries).

We proposed a new formalism, the Allsome Sequence Bloom Trees [28]. It improves upon Sequence Bloom Trees in terms of construction time (by 50%) and query time (by 40-85%), and also permits dataset-vs-dataset searches. The method has been tested by indexing a subset of 2,652 RNA-seq human experiments from the Sequence Read Archive. Allsome Sequence Bloom Trees pave the way towards "Google" searches of petabytes of sequencing data.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

- ANR ASTER: ASTER is a national project that aims at developing algorithms and software for analyzing third-generation sequencing data, and more specifically RNA sequencing. BONSAI is the principal investigator in this ANR. Other partners are Erable (LBBE in Lyon) and two sequencing and analysis platforms that have been very active in the MinION Access Program (Genoscope and Institut Pasteur de Lille).
- PIA France Génomique: National funding from “Investissements d’Avenir” (call *Infrastructures en Biologie-Santé*). France Génomique is a shared infrastructure, whose goal is to support sequencing, genotyping and associated computational analysis, and increases French capacities in genome and bioinformatics data analysis. It gathers 9 sequencing and 8 bioinformatics platforms. Within this consortium, we are responsible for the workpackage devoted to the computational analysis of sRNA-seq data, in coordination with the bioinformatics platform of Génopole Toulouse-Midi-Pyrénées.

8.1.2. ADT

- ADT Vidjil (2015–2017): The purpose of this ADT was to strengthen Vidjil development and to ensure a better diffusion of the software by easing its installation, administration and usability. This enabled the software to be well suited for a daily clinical use. Vidjil is now used in routine practice by seven European hospitals (France, Germany, Italy and Czech Republic). Hospitals from the United Kingdom and the Japan are currently assessing Vidjil and may do their clinical routine practice with the software in a near future.
- ADT SeedLib (2017–2019): The SeedLib ADT aims to consolidate existing software developments in Bonsai, into an existing and well-engineered framework. Bonsai has published several new results on spaced seeds and developed several tools that integrate custom implementations of spaced seeds. In parallel, the GATB project is a C++ software library that facilitates the development of next-generation sequencing analysis tools. It is currently maintained by a collaboration between the GenScale team at Inria Rennes and the Bonsai team. Many users from other institutions (including the Erable team at Inria Rhones-Alpes) actively develop tools using GATB. The core object in GATB is k-mers, which can be seen as the predecessor of spaced seeds. The goal of this ADT is to integrate existing space seeds formalisms into GATB, therefore further expanding the features offered by the library, and at the same time provide visibility for tools and results in the Bonsai team.

8.2. European Initiatives

8.2.1. Collaborations in European Programs, Except FP7 & H2020

- International ANR RNAlands (2014-2018): National funding from the French Agency Research (call *International call*). Our objective is the fast and efficient sampling of structures in RNA Folding Landscapes. The project gathers three partners: Amib from Inria Saclay, the Theoretical Biochemistry Group from Universität Wien and BONSAI.
- Interreg Va (France-Wallonie-Vlaanderen): Portfolio "SmartBioControl", including 5 constitutive projects and 25 partners working together towards sustainable agriculture.

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

- *Astrid Lindgrens Hospital, Stockholm University:* Collaboration with Anna Nilsson and Shanie Saghafian-Hedengren on RNA sequencing of stromal cells (pilot study done in 2017).
- *Childhood Leukaemia Investigation Prague (CLIP), Department of Pediatric Hematology/Oncology, 2nd Faculty of Medicine, Charles University, Prague, Czech Republic:* Collaboration with Michaela Kotrová and Eva Fronkova on leukemia diagnosis and follow-up.
- *CWI Amsterdam:* Collaboration with Alexander Schoenhuth on data structures for genomic data.
- *Department of Statitics, North Carolina State University:* Collaboration with Donald E. K. Martin on spaced seeds coverage [21].
- *Département des Sciences de la Vie, Faculté des Sciences de Liège:* Collaboration with Denis Beaurain on nonribosomal peptides.
- *Gembloux Agro-Bio Tech, Université de Liège:* Collaboration with Philippe Jacques on nonribosomal peptides.
- *Institute of Biosciences and Bioresources, Bari:* Collaboration with Nunzia Scotti on the assembly of plant mitochondrial genomes.
- *Medvedev lab, The Pennsylvania State University:* Collaboration with Paul Medvedev on algorithms and data structures for genomic data, e.g. the Allsome Sequence Bloom Trees.
- *Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark:* Collaboration with Tilmann Weber on nonribosomal peptides.

- *Proteome Informatics Group, Swiss Institute of Bioinformatics:* Collaboration with Frédérique Lisacek on nonribosomal peptides.
- *School of Social and Community Medicine, University of Bristol:* Collaboration with John Moppett and Stephanie Wakeman on leukemia diagnosis follow-up.
- *Theoretical Biochemistry Group, Universität Wien:* Collaboration with Andrea Tanzer and Ronny Lorenz on RNA folding and RNA kinetics.

8.3.2. Participation in Other International Programs

- Participation in the EuroClonality-NGS consortium. This consortium aims at standardizing the study of immune repertoire, clonality and minimal residual disease in leukemia at the european level. We are part of the bioinformatics workgroup led by Nikos Darzentas (CEITEC, Brno, Czech Republic). Withing this consortium, we participated to a lead opinion paper on immunohematology [20].

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

- JOBIM 2017, national conference in bioinformatics (J.-S. Varré for the organizing committee, H. Touzet for the program committee).
- SeqBio 2017, national workshop on stringology, combinatorics and bioinformatics (M. Salson).
- From RNA-Seq data to bioinformatics analysis using Nanopore sequencers 2017 (H. Touzet).

9.1.1.2. Member of the Organizing Committees

- JOBIM 2017, national conference on bioinformatics (A. Béliard, S. Janot, P. Marijon, L. Noé, P. Pericard, M. Pupin, T. Rocher, C. Saad, M. Salson).
- SeqBio 2017, national workshop on stringology, combinatorics and bioinformatics (R. Chikhi, S. Janot, L. Noé, H. Touzet).

9.1.2. Scientific Events Selection

9.1.2.1. Member of the Conference Program Committees

- WABI 2017 (M. Salson, H. Touzet).
- BCB 2017 (R. Chikhi).
- HiTSEQ 2017 (R. Chikhi).
- CSR 2017 (H. Touzet).
- JOBIM 2017 (S. Blanquart, R. Chikhi).

9.1.2.2. Reviewer

- IWOCA 2017 (M. Salson).
- WALCOM 2017 (M. Giraud, M. Salson),
- LAGOS 2017 (R. Chikhi).
- WG 2017 (R. Chikhi).

9.1.3. Journal

9.1.3.1. Reviewer - Reviewing Activities

- Bioinformatics (M. Salson, R. Chikhi, L. Noé).
- Discrete Applied Mathematics (M. Salson).

- NAR (R. Chikhi).
- BMC Bioinformatics (R. Chikhi, L. Noé).
- PeerJ (L. Noé, M. Giraud).

9.1.4. Research Administration

- Member of the CUB for Inria Lille (S. Blanquart).
- Member of the CDT for Inria Lille (M. Pupin).
- Member of the CUMI for Inria Lille (M. Salson).
- Member of the national scientific committee of INS2I–CNRS (H. Touzet).
- Member of the scientific committee of MBIA – INRA (H. Touzet).
- Head of the national CNRS network GDR Bioinformatique moléculaire (http://www.gdr-bim.cnrs.fr, H. Touzet).
- Co-head of the Lille Bioinformatics platform, bilille (H. Touzet).
- Member of the CRIStAL Laboratory council (H. Touzet).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Teaching in computer science:

- License: L. Noé, *Networks*, 42h, L3 Computer science, Univ. Lille 1.
- License: L. Noé, *Programming (Python)*, 54h, L3 Computer science' S3H, Univ. Lille 1.
- License: L. Noé, *Coding and information theory*, 36h, L2 Computer science, Univ. Lille 1.
- License: L. Noé, *Functional Programming*, 30h, L2 Computer science, Univ. Lille 1.
- License: J.-S. Varré, *Object oriented programming*, 36h, L2 Computer Science, Univ. Lille 1.
- License: J.-S. Varré, *Algorithms and data structures*, 50h, L2 Computer science, Univ. Lille 1.
- License: J.-S. Varré, *System*, 84h, L3 Computer science, Univ. Lille 1.
- Master: J.-S. Varré, M. Salson, *Software project*, 40h, M1 Computer science, Univ. Lille 1.
- License: P. Marijon *Databases*, 36h, L3 Computer science, Univ. Lille 1.
- License: S. Janot, *Introduction to programming (C)*, 50h, L3 Polytech'Lille, Univ. Lille 1.
- License: S. Janot, *Databases*, 30h, L3 Polytech'Lille, Univ. Lille 1.
- Master: S. Janot, *Databases*, 12h, M1 Polytech'Lille, Univ. Lille 1.
- Master: S. Janot, *Logic and Semantic Web*, 80h, M1 Polytech'Lille, Univ. Lille 1.
- Licence: M. Pupin, *Programming (Python)*, 78h, L1 Sciences, Univ. Lille 1.
- Licence: M. Pupin, *occupational integration*, 30h, L3 computer science, Univ. Lille 1.
- Master: : M. Pupin, *Programming (JAVA)*, 24h, M1 "Mathématiques et finance", Univ. Lille 1.
- License: C. Saad, *Algorithmics and programming*, 28h, L3 Polytech'Lille, Univ. Lille 1.
- License: C. Saad, *Databases*, 36h, L3 Polytech'Lille, Univ. Lille 1.
- License: C. Saad, *Data structures*, 18h, L3 Polytech'Lille, Univ. Lille 1.
- License: C. Saad, *Introduction to programming (C)*, 34h, L3 Polytech'Lille, Univ. Lille 1.
- License: C. Saad, *Programming (C)*, 22h, L3 Polytech'Lille, Univ. Lille 1.
- License: M. Salson, *Programming (Python)*, 42h, L1 Sciences, Univ. Lille 1.
- License: M. Salson, *Coding and information theory*, 63h, L2 Computer science, Univ. Lille 1.

Teaching in bioinformatics:

- License: S. Blanquart, R. Chikhi, M. Giraud, *Bioinformatics*, 40h, L3 Computer Science, Univ. Lille 1.
- Master: S. Blanquart, *Algorithms and applications in bioinformatics*, 24h, M1 Computer Science, Univ. Lille 1.
- Master: S. Blanquart, *Methods in phylogenetics*, 4h, M2 Biodiversité Evolution Ecologie, Univ. Lille 1.
- Master: L. Noé, *Bioinformatics*, 40h, M1 Biotechnologies, Univ. Lille 1.
- Master: M. Pupin, *Bioinformatics*, 34h, M1 "Biologie-Santé", Univ. Lille 1.
- Master: M. Salson, *Algorithms for life sciences*, 20h, M2 Complex models, algorithms and data, Univ. Lille 1.

Teaching in skeptical thinking:

- License: M. Giraud, 35h, L3 Computer science, Univ. Lille 1.
- Master: M. Salson, *Skeptical thinking*, 27h, M2 Journalist and Scientist, ESJ, Univ. Lille 1.

Formation for academics:

- Bilille permanent training: S. Blanquart (*Phylogenetics*, 28h), M. Pupin (*Gene prediction and protein annotation*, 14h), J.S. Varré (*Gene prediction and protein annotation*, 14h), C. Saad (*Variants*, 13h), R. Chikhi (*De novo assembly* and *Metagenomics de novo assembly*, 8h), H. Touzet (*Metagenomics, 3h)*, L. Noé (*DNAmapping*, 3h), M. Salson (*RNA-seq analysis*, 1h).
- Workshop on Genomics (Czech Republic): R. Chikhi *De novo assembly*, 8h, young researchers.
- Roscoff summer school in Bioinformatics: R. Chikhi *De novo assembly*, 8h, engineers and researchers.
- Software training (Granada, Spain and Rennes, France): R. Chikhi *Using the GATB library*, 2x8h, engineers and researchers.

9.2.2. Teaching administration

- Head of the licence semester "Computer Science – S3 Harmonisation (S3H)", Univ. Lille 1 (L. Noé).
- Member of faculty council (M. Pupin, J.-S. Varré).
- Head of the 3rd year of licence of computer science, Univ. Lille 1 (J.-S. Varré).
- Head of the GIS department (Software Engineering and Statistics) of Polytech'Lille (S. Janot).
- Head of the computer science modules in the 1st year of Licence, Univ. Lille 1 (M. Pupin).

9.2.3. Supervision

- PhD: L. Siegwald, Solutions d'amélioration des études de métagénomique ciblée, 2017/03/23, H. Touzet, Y. Lemoine.
- PhD: P. Pericard, Algorithmes pour la reconstruction de séquences de marqueurs conservés dans des données de métagénomique, 2017/10/27, H. Touzet, S. Blanquart.
- PhD in progress: T. Rocher, Indexing VDJ recombinations in lymphocytes for leukemia follow-up, 2014/11/01, M. Giraud, M. Salson.
- PhD in progress: C. Saad, Caractérisation des erreurs de séquençage non aléatoires, application aux mosaïques et tumeurs hétérogènes, 2014/10/01, M.-P. Buisine, H. Touzet, J. Leclerc, L. Noé, M. Figeac.
- PhD in progress: Q. Bonenfant, Algorithmes pour l'analyse de séquençage ARN troisième génération, 2017/11/15, L. Noé, H. Touzet.
- PhD in progress: P. Marijon, Analyse de graphes d'assemblage issus du séquençage ADN troisième génération, 2016, R. Chikhi, J-S. Varré.

9.2.4. Juries

- H. Touzet was member of the PhD juries of Antoine Limasset (Université Rennes 1), Anna Kuosmanen (University of Helsinki), Jean-Pierre Glouzon (University of Sherbrooke), Guillaume Madeleine (Université Lille 1) and Jananan Pathmanathan (UPMC).
- R. Chikhi was member of the PhD juries of Kamil Salikhov (Université de Paris-Est) and Guillaume Holley (Bielefeld University).
- M. Salson was member of the PhD jury of Alice Héliou (Université Paris-Sud).
- H. Touzet was member of hiring committees (professors) at Université de Nantes and Université de Strasbourg.

9.3. Popularization

- *Recruiting girls to computer science* M. Pupin and T. Rocher are members of the collective *Informatique au féminin* from University of Lille, which was launched three years ago and whose goal is to organize computer science initiatives that reach teenage girls and female students. Among other actions, they were fully involved in the event *L codent, L créent* (she codes, she creates). This action aims to teach code to schoolgirls (13-15 years old), before they amass prejudices against computer science. 35 teenage girls were supervised by 9 female graduate computer science students, to create a proximity link between the young women. To emphasize the fact that coding is a creative and innovative pursuit, we chose to teach *Processing*, a programming language built for visual arts. After eight sessions of creative coding, a public exhibition was organized at the University with inspirational testimonies of women working in the field of computer science. This experiment has been presented at the conference PyParis2017 [30] and at the ACM conference womENcourage2017 [31].
- The team participates to dissemination actions for high school students and high school teachers on a regular basis: multiple presentations on bioinformatics and research in bioinformatics with our dedicated "genome puzzles", practical session at "Day for Programming and Algorithmic Teaching", presentations at "Salon de l'étudiant", visit of high school students in the team (M. Giraud, M. Salson, J.-S. Varré)
- *Explaining big data to a general audience*: H. Touzet was part of the editorial committee and author for the book *"Les big data à découvert"*(368 pages).

10. Bibliography

Major publications by the team in recent years

[1] A. ABDO, S. CABOCHE, V. LECLÈRE, P. JACQUES, M. PUPIN. *A new fingerprint to predict nonribosomal peptides activity*, in "Journal of Computer-Aided Molecular Design", October 2012, vol. 26, n^{o} 10, pp. 1187-94 [*DOI :* 10.1007/s10822-012-9608-4], http://hal.inria.fr/hal-00750002

[2] A. ABDO, V. LECLÈRE, P. JACQUES, N. SALIM, M. PUPIN. *Prediction of new bioactive molecules using a bayesian belief network*, in "Journal of Chemical Information and Modeling", January 2014, vol. 54, n^{o} 1, pp. 30-36 [*DOI :* 10.1021/ci4004909], https://hal.archives-ouvertes.fr/hal-01090611

[3] R. CHIKHI, A. LIMASSET, P. MEDVEDEV. *Compacting de Bruijn graphs from sequencing data quickly and in low memory*, in "Bioinformatics", November 2016, vol. 32, n^{o} 12, pp. i201 - i208 [*DOI :* 10.1093/bioinformatics/btw279], https://hal.archives-ouvertes.fr/hal-01395704

[4] Y. DUFRESNE, L. NOÉ, V. LECLÈRE, M. PUPIN. *Smiles2Monomers: a link between chemical and biological structures for polymers*, in "Journal of Cheminformatics", December 2015 [*DOI :* 10.1186/S13321-015-0111-5], https://hal.inria.fr/hal-01250619

[5] Y. FERRET, A. CAILLAULT, S. SEBDA, M. DUEZ, N. GRARDEL, N. DUPLOYEZ, C. VILLENET, M. FIGEAC, C. PREUDHOMME, M. SALSON, M. GIRAUD. *Multi-loci diagnosis of acute lymphoblastic leukaemia with high-throughput sequencing and bioinformatics analysis*, in "British Journal of Haematology", 2016, bjh.13981 p. [*DOI :* 10.1111/BJH.13981], https://hal.archives-ouvertes.fr/hal-01279160

[6] A. FLISSI, Y. DUFRESNE, J. MICHALIK, L. TONON, S. JANOT, L. NOÉ, P. JACQUES, V. LECLÈRE, M. PUPIN. *Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing*, in "Nucleic Acids Research", 2015 [*DOI :* 10.1093/NAR/GKV1143], https://hal.archives-ouvertes.fr/hal-01235996

[7] M. FRITH, L. NOÉ. *Improved search heuristics find 20 000 new alignments between human and mouse genomes*, in "Nucleic Acids Research", February 2014, vol. 42, n^{o} 7, e59 p. [*DOI :* 10.1093/NAR/GKU104], https://hal.inria.fr/hal-00958207

[8] R. GIEGERICH, H. TOUZET. *Modeling dynamic programming problems over sequences and trees with inverse coupled rewrite systems*, in "Algorithms", 2014, vol. 7, pp. 62 - 144 [*DOI :* 10.3390/A7010062], https://hal.archives-ouvertes.fr/hal-01084318

[9] M. GIRAUD, M. SALSON, M. DUEZ, C. VILLENET, S. QUIEF, A. CAILLAULT, N. GRARDEL, C. ROUMIER, C. PREUDHOMME, M. FIGEAC. *Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing*, in "BMC Genomics", 2014, vol. 15, n^{o} 1, 409 p. [*DOI :* 10.1186/1471-2164-15-409], https://hal.archives-ouvertes.fr/hal-01009173

[10] E. KOPYLOVA, L. NOÉ, H. TOUZET. *SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data*, in "Bioinformatics", October 2012, pp. 1-10 [*DOI :* 10.1093/BIOINFORMATICS/BTS611], http://hal.inria.fr/hal-00748990

[11] M. LÉONARD, L. MOUCHARD, M. SALSON. *On the number of elements to reorder when updating a suffix array*, in "Journal of Discrete Algorithms", February 2012, vol. 11, pp. 87-99 [*DOI :* 10.1016/J.JDA.2011.01.002], http://hal.inria.fr/inria-00636066

[12] D. E. K. MARTIN, L. NOÉ. *Faster exact distributions of pattern statistics through sequential elimination of states*, in "Annals of the Institute of Statistical Mathematics", September 2015 [*DOI :* 10.1007/S10463-015-0540-Y], https://hal.inria.fr/hal-01237045

[13] L. NOÉ, D. E. K. MARTIN. *A coverage criterion for spaced seeds and its applications to support vector machine string kernels and k-mer distances*, in "Journal of Computational Biology", November 2014, vol. 21, n^{o} 12, 28 p. [*DOI :* 10.1089/CMB.2014.0173], https://hal.inria.fr/hal-01083204

[14] A. PERRIN, J.-S. VARRÉ, S. BLANQUART, A. OUANGRAOUA. *ProCARs: progressive reconstruction of ancestral gene orders*, in "BMC Genomics", 2015, vol. 16, n^{o} Suppl 5, S6 p. [*DOI :* 10.1186/1471-2164-16-S5-S6], https://hal.inria.fr/hal-01217311

[15] M. PUPIN, Q. ESMAEEL, A. FLISSI, Y. DUFRESNE, P. JACQUES, V. LECLÈRE. *Norine: a powerful resource for novel nonribosomal peptide discovery*, in "Synthetic and Systems Biotechnology", December 2015 [*DOI :* 10.1016/J.SYNBIO.2015.11.001], https://hal.inria.fr/hal-01250614

[16] A. SAFFARIAN, M. GIRAUD, A. DE MONTE, H. TOUZET. *RNA locally optimal secondary structures*, in "Journal of Computational Biology", 2012, vol. 19, n^{o} 10, pp. 1120-1133 [*DOI :* 10.1089/CMB.2010.0178], http://hal.inria.fr/hal-00756249

[17] A. SAFFARIAN, M. GIRAUD, H. TOUZET. *Modeling alternate RNA structures in genomic sequences*, in "Journal of Computational Biology", February 2015, vol. 22, n^{o} 3, pp. 190-204, https://hal.archives-ouvertes.fr/hal-01228130

Publications of the year

Articles in International Peer-Reviewed Journals

[18] J. AUDOUX, M. SALSON, C. F. GROSSET, S. BEAUMEUNIER, J.-M. HOLDER, T. COMMES, N. PHILIPPE. *SimBA: A methodology and tools for evaluating the performance of RNA-Seq bioinformatic pipelines*, in "BMC Bioinformatics", September 2017, vol. 18, n^{o} 1, 428 p. [*DOI :* 10.1186/S12859-017-1831-5], http://www.hal.inserm.fr/inserm-01612738

[19] Q. ESMAEEL, M. PUPIN, P. JACQUES, V. LECLÈRE. *Nonribosomal peptides and polyketides of Burkholderia: new compounds potentially implicated in biocontrol and pharmaceuticals*, in "Environmental Science and Pollution Research", May 2017 [*DOI :* 10.1007/S11356-017-9166-3], https://hal.archives-ouvertes.fr/hal-01548616

[20] A. W. LANGERAK, M. BRÜGGEMANN, F. DAVI, N. DARZENTAS, J. J. M. VAN DONGEN, D. GONZALEZ, G. CAZZANIGA, V. GIUDICELLI, M.-P. LEFRANC, M. GIRAUD, E. A. MACINTYRE, M. HUMMEL, C. POTT, P. J. T. A. GROENEN, K. STAMATOPOULOS. *High-Throughput Immunogenetics for Clinical and Research Applications in Immunohematology: Potential and Challenges*, in "Journal of Immunology", April 2017, 1602050 p. [*DOI :* 10.4049/JIMMUNOL.1602050], https://hal.archives-ouvertes.fr/hal-01516289

[21] D. E. K. MARTIN, L. NOÉ. *Faster exact distributions of pattern statistics through sequential elimination of states*, in "Annals of the Institute of Statistical Mathematics", February 2017, vol. 69, n^{o} 1, pp. 231–248 [*DOI :* 10.1007/S10463-015-0540-Y], https://hal.inria.fr/hal-01237045

[22] L. NOÉ. *Best hits of 11110110111: model-free selection and parameter-free sensitivity calculation of spaced seeds*, in "Algorithms for Molecular Biology", February 2017, vol. 12, n^{o} 1 [*DOI :* 10.1186/S13015-017-0092-1], https://hal.inria.fr/hal-01467970

[23] P. PERICARD, Y. DUFRESNE, L. COUDERC, S. BLANQUART, H. TOUZET. *MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes*, in "Bioinformatics", October 2017 [*DOI :* 10.1093/BIOINFORMATICS/BTX644], https://hal.inria.fr/hal-01646297

[24] M. SALSON, M. GIRAUD, A. CAILLAULT, N. GRARDEL, N. DUPLOYEZ, Y. FERRET, M. DUEZ, R. HERBERT, T. ROCHER, S. SEBDA, S. QUIEF, C. VILLENET, M. FIGEAC, C. PREUDHOMME. *High-throughput sequencing in acute lymphoblastic leukemia: Follow-up of minimal residual disease and emergence of new clones*, in "Leukemia Research", 2017, vol. 53, pp. 1-7 [*DOI :* 10.1016/J.LEUKRES.2016.11.009], https://hal.archives-ouvertes.fr/hal-01404817

[25] A. SCZYRBA, P. HOFMANN, P. BELMANN, D. KOSLICKI, S. JANSSEN, J. DRÖGE, I. GREGOR, S. MAJDA, J. FIEDLER, E. DAHMS, A. BREMGES, A. FRITZ, R. GARRIDO-OTER, T. S. JØRGENSEN, N. SHAPIRO, P. D. BLOOD, A. GUREVICH, Y. BAI, D. TURAEV, M. Z. DEMAERE, R. CHIKHI, N. NAGARAJAN, C. QUINCE, L. H. HANSEN, S. J. SØRENSEN, B. K. H. CHIA, B. DENIS, J. L. FROULA, Z. WANG, R. EGAN, D. DON KANG, J. J. COOK, C. DELTEL, M. BECKSTETTE, C. LEMAITRE, P. PETERLONGO, G. RIZK, D. LAVENIER, Y.-W. WU, S. W. SINGER, C. JAIN, M. STROUS, H. KLINGENBERG, P. MEINICKE, M. D. BARTON, T. LINGNER, H.-H. LIN, Y.-C. LIAO, G. G. Z. SILVA, D. A. CUEVAS, R. A. EDWARDS, S. SAHA, V. C. PIRO, B. Y. RENARD, M. POP, H.-P. KLENK, M. GÖKER, N. C. KYRPIDES, T. WOYKE, J. A. VORHOLT, P. SCHULZE-LEFERT, E. M. RUBIN, A. E. DARLING, T. RATTEI, A. C. MCHARDY. *Critical Assessment of Metagenome Interpretation – a benchmark of computational metagenomics software*, in "Nature Methods", October 2017, vol. 14, n° 11, pp. 1063 - 1071 [*DOI :* 10.1038/NMETH.4458], https://hal.archives-ouvertes.fr/hal-01633525

International Conferences with Proceedings

[26] A. LIMASSET, G. RIZK, R. CHIKHI, P. PETERLONGO. *Fast and scalable minimal perfect hashing for massive key sets*, in "16th International Symposium on Experimental Algorithms", London, United Kingdom, June 2017, vol. 11, pp. 1 - 11, https://arxiv.org/abs/1702.03154 , https://hal.inria.fr/hal-01566246

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

[28] C. SUN, R. S. HARRIS, R. CHIKHI, P. MEDVEDEV. *AllSome Sequence Bloom Trees*, in "RECOMB 2017 - 21st Annual International Conference on Research in Computational Molecular Biology", Hong Kong, China, May 2017 [*DOI :* 10.1007/978-3-319-56970-3_17], https://hal.inria.fr/hal-01575350

National Conferences with Proceedings

[29] *Best Paper*
P. PERICARD, Y. DUFRESNE, S. BLANQUART, H. TOUZET. *Reconstruction of full-length 16S rRNA sequences for taxonomic assignment inmetagenomics*, in "JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques", Lille, France, July 2017, https://hal.inria.fr/hal-01574629.

Scientific Popularization

[30] M. PUPIN, P. MARQUET, Y. SECQ. *How to make teenage girls love coding using Python and the visual arts orienting language Processing ?*, in "PyParis2017", Paris, La Défense, France, Systematic Paris Region, June 2017, https://hal.inria.fr/hal-01552487

[31] T. ROCHER, M. PUPIN, P. MARQUET, Y. SECQ. *How to make teenage girls love coding ?*, in "womENcourage2017 - 4th ACM Europe Celebration of Women in Computing", Barcelona, Spain, ACM, September 2017, https://hal.inria.fr/hal-01552490

Other Publications

[32] P. MARIJON, J.-S. VARRÉ, R. CHIKHI. *Debugging long-read genome and metagenome assemblies using string graph analysis*, July 2017, JOBIM 2017- Journées Ouvertes en Biologie, Informatique et Mathématiques, Poster, https://hal.inria.fr/hal-01574824

[33] C. SAAD, L. NOÉ, H. RICHARD, J. LECLERC, M.-P. BUISINE, H. TOUZET, M. FIGEAC. *DiNAMO: Exact method for degenerate IUPAC motifs discovery, characterization of sequence-specific errors*, July 2017, JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques, Poster, https://hal.inria.fr/hal-01574630