IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2018

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, end of the Project-Team: 2018 December 31

Keywords:

Computer Science and Digital Science:

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

Other Research Topics and Application Domains:

B1.1.5. - Immunology
B1.1.6. - Evolutionnary biology
B1.1.7. - Bioinformatics
B1.1.11. - Plant Biology
B2.2.3. - Cancer

1. Team, Visitors, External Collaborators

Research Scientists

Hélène Touzet [Team leader, CNRS, Senior Researcher, HDR]
Samuel Blanquart [Inria, Researcher, until Apr 2018]
Rayan Chikhi [CNRS, Researcher]

Faculty Members

Stéphane Janot [Université de Lille, Associate Professor]
Laurent Noé [Université de Lille, Associate Professor]
Maude Pupin [Université de Lille, Associate Professor, HDR]
Mikaël Salson [Université de Lille, Associate Professor]
Jean-Stéphane Varré [Université de Lille, Professor, HDR]

Post-Doctoral Fellow

Aymeric Antoine-Lorquin [Inria]

PhD Students

Quentin Bonenfant [CNRS]
Pierre Marijon [Inria]
Chadi Saad [Université de Lille, until Aug 2018]

Technical staff

Areski Flissi [CNRS]
Ryan Herbert [Inria, until Mar 2018]
Mael Kerbiriou [Inria]

Interns

Quentin Charret [Inria, from Feb 2018 until Jul 2018]
Clementine Campart [CNRS, from March 2018]

Administrative Assistant

Amelie 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. Highlights of the Year

4.1. Highlights of the Year

The team was actively involved in the organization of the international conference RECOMB in Paris (April 2018), that was attended by more than 800 people.

4.1.1. Awards

First place at the metagenomics assembly challenge organized by the company Mosaic DNANexus: https://www.businesswire.com/news/home/20180620005408/en/DNAnexus-Powered-Mosaic-Microbiome-Platform-Announces-Winners-Community

5. New Software and Platforms

5.1. 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

5.2. 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

5.3. 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

5.4. 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

6. New Results

6.1. Exploration of transcriptomes

In 2016 we produced a method called CG-Alcode able to compare transcripts repertoires of a given pair of orthologous genes. We applied our method to compare human and mouse transcriptomes. This year, in collaboration with C.Belleannée (DYLISS, Inria Rennes) we explored the comparison of multiple species. We inspected human, mouse and dog transcriptomes. We thus were able to predict a large number of putative transcripts in both human, mouse and dog based on known transcripts. Those results allow to investigate which functional sites are conserved and which genes have the same set of transcripts (known or putative).

6.2. Modeling of alternative transcripts with long reads

In the context of transcriptomic analyses based on third generation sequencing data (ONT), we started to explore the following problem : given a transcriptomic experiment, a gene of interest, select reads related to the given gene and find exon junctions. As we have done in the CG-Alcode project, we aim to model the gene as an alphabet of exonic blocks, each transcript being a word over this alphabet. This work takes place in the context of ANR ASTER for which we deal with mouse transcriptomic data in brain and liver. Built models will allow to query human genes to discover putative transcripts.

6.3. Read against read comparison for Nanopore data

In the team, we developed two years ago seeds with errors, which allow to find all common approximate patterns with a limited number of errors. The idea behind these seeds, called 01^*0 seeds, is to divide the sequence in blocks so that the distribution of errors is no longer random. This year, we have used these seeds in the context of long reads analysis. With this data, reads against reads comparison suffers from a high loss of sensitivity, because the single *read error-rate* is already high. Our application case is the detection of adapter sequences in ONT sequencing. We have shown that the use of these seeds instead of exact k-mers allowed a more accurate reconstruction of the sequences of the adapters. The method takes two steps: first the identification of k-mers potentially composing the adapter using a counting approach that takes into account errors in the read, and then the reconstruction of the complete sequence of the adapter with a greedy algorithm. Our results show that the seeds with errors allow to obtain accurate consensus sequences for more 80% of the samples, compared to 40% with the usual k-mer approach. This work was done within the ANR ASTER during the first year of the thesis of Quentin Bonenfant and was presented at the national workshop Seqbio 2018.

6.4. Annotation of the OC43 coronavirus genome

OC43 coronavirus is recognized as frequent cause of respiratory infection. We have conducted a bioinformatics study of 8 coronavirus genomes collected from patients at Lille hospital : gene annotation, phylogenetic analysis and amino acids substitutions. Several genotypes (B, E, F and G) were identified and two clusters of patients were defined from chronological data and phylogenetic trees based on the genomic sequences,. Analyses of amino acids substitutions of the S protein sequences identify substitutions specific of genotype F strains circulating among French people. This work is a collaboration with Anne Goffard (CHRU Lille and CIIL).

6.5. Small RNAs catalog in oilseed rape

Polyploidy – and notably allopolyploidy that involves interspecific hybridization – has played a major role in the evolution of plants, partly because this process is often associated with genomic structure and expression changes. Homeologous exchanges (HE) – i.e. between the constituent subgenomes– have been demonstrated to be frequent in allopolyploids and could be involved in the origin and maintenance of polyploids. While their influence on gene content has poorly been studied until recently, little is known about their impact on gene expression. Together with K. Alix (Inra Moulon), we have analyzed the impact of HEs that have been characterized in resynthesized oilseed rapes, on the repertoire of micro RNAs. Our main objective was to assess the relations that could exist between structural variation and modifications of gene expression through changes in miRNA regulation. The analysis was based on the small RNA-seq catalog obtained with the bioinformatic tool miRkwood, developed in BONSAI. We have built a microRNA database for the diploid subgenomes AA from *Brassica rapa* and CC for *Brassica oleracea* that correspond to the progenitors of the resynthesized *Brassica napus* allotetraploids (AACC). Integrating miRNA prediction and genomic location of HEs allowed us to infer relationships between microRNA restructuring and non-additivity of gene expression in polyploid hybrids.

6.6. Identifying systematic sequencing errors

Discovering over-represented approximate motifs in DNA sequences is an essential part of bioinformatics, which has been studied extensively. However, it remains a difficult challenge, especially with the huge quantity of data generated by high throughput sequencing technologies. We have developed an exact discriminative method for IUPAC motifs discovery in large sets of DNA sequences. The approach uses mutual information (MI) as an objective function to search for over-represented degenerate motifs in a lattice [7].

The algorithm was applied to the problem of *Sequence-Specific Errors*. Next Generation Sequencing, and further Single-Molecule Sequencing technologies are known to produce a highly variable error rate. A common method to overcome these sequencing errors is to increase the *coverage*. However, Sequence-Specific Errors are recurrent errors that depend on the upstream nucleotidic context, and can thus be confused with true genomic variations when the read coverage increases. Our algorithm was able to find motifs associated to sequencing errors and therefore to improve variant calling. This method has also tested on ChIP-seq datasets, and compared with five state-of-the art methods, where it was experimentally shown to perform as well as the best one, while be resistant to down-sampling.

This work was done during the thesis of Chadi Saad, and as a collaboration with Martin Figeac (Univ. Lille - Plateau de génomique fonctionnelle et structurale), Julie Leclerc and Marie-Pierre Buisine (CHRU de Lille - JPARC), and Hugues Richard (Sorbonne Université - Laboratory Computational and Quantitative Biology).

6.7. Indexing labelled sequences

We designed a compressed full-text index structure able to index a whole text with labels attached to every letter in the text [6]. This work will be applied to DNA sequences and more precisely V(D)J recombinations which are complex genomic rearrangements occurring in lymphocytes. The index will be used to index labelled V(D)J recombinations, which are labelled with their V, D and J gene. As the index we conceived is scalable, we will index V(D)J recombinations from thousands of samples and give access to this data through the Vidjil platform.

6.8. Tree representations

We found an intriguing duality between two well-known representations of trees [12]. This work concerns data structures and succinct tree representations. The Balanced Parenthesis representation of trees consists of encoding the structure of any tree using a series of opening and closing parentheses. The DFUDS representation is similar, but differs in how each node is encoded (also using parentheses). By relating both BP and DFUDS representations, we obtained improvements for a basic fundamental problem: the Minimum Length Interval Query problem. We also reported unnoted commonalities in recent solution to the Range Minimum Query problem.

6.9. Co-linear chaining on graphs

We reported the first algorithm that perform co-linear chaining between a sequence and a directed acyclic graph (DAG) [9]. This work concerns dynamic programming algorithms and sequencing alignment. The problem of co-linear chaining is a classical bioinformatics problem, which has immediate application to sequence alignment, as it is used as a filter remove spurious alignment seeds. Co-linear chaining is typically solved using a simple dynamic programming algorithm. Yet, representations of genomes using graphs instead of sequences have recently became an active research topic. As a results, the problem of aligning a sequence to a sequence graph merits consideration. This work provides the first step towards tackling practical sequence-to-graph alignment instances, by first considering the case when the graph is a DAG. We designed a $O(k|E|log|V|)$ algorithm to solve co-linear chaining on DAGs, which matches the optimal solution for the classical sequence variant, i.e. when the graph is a path.

6.10. Representations of de Bruijn graphs

We designed the first practical data structure for representing large de Bruijn graphs, which supports insertions and deletions of nodes [3]. This work concerns *de novo* assembly and several other k-mer-related bioinformatics problems. The representation of de Bruijn graphs is a transversal bioinformatics question that has enjoyed recent applications in genome, metagenome and transcriptome assembly and quantification. To this date, efficient data structures were essentially static. In this work we provided an implementation of a dynamic data structure that combines perfect hashing, Karp-Rabin hashing, and forests. Practical tests show that this structure is highly competitive with the state of the art.

6.11. Readability of overlap graphs

We report further progress on the study of a theoretical parameter of graph named *readability* [8]. This work concerns graph theory mainly. The readability parameter measures the minimal length of strings that would be needed in order to label a graph such that it is an overlap graph over a set of strings of that length. So far, recent works on readability have not elucidated many aspects related to this parameter: the complexity of computing it is open, and it is not even known whether the corresponding decision problem is in NP. The only upper bound known for this parameter is exponential. This work focuses on certain graph families: bipartite chain graphs, grids, induced subgraphs of grids, and provides a characterization of bipartite graphs of readability 2.

6.12. Nonribosomal peptides

Norine is a comprehensive public database for non-ribosomal peptides developed by the team for more than 10 years. The Norine database quality has been enhanced through a semi-automatic curation process of data. Particularly, more than 500 SMILES annotations have been added or updated. This allowed us to check and correct the monomeric graphs, i.e. a 2D representation of the monomeric composition of the NRPs, thanks to dedicated tools like Smiles2Monomers. This update was done in collaboration with members of the Proteome Informatics Group from SIB (Swiss Institute of Bioinformatics). New annotations on monoisotopic mass and molecular formulas have also been added. The Norine interface was improved and new features are available, such as the possibility to access the complete change history of each entry. To encourage new submissions of NRPs, authors of new NRPs are now visible as contributors on Norine home page. Finally, we published this year, in the field of biocontrol (a contraction of “biological control”), a paper on bioinformatic tools for the discovery of new lipopeptides [5], essentially based on the Norine platform.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. ANR

- ANR Transipedia: The purpose of Transipedia is to provide means of identifying relevant transcriptomic events within thousands of RNA sequencing experiments. This project will be achieved in collaboration with I2BC (principal investigator) in Paris Saclay and IRMB in Montpellier.
- 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 work package devoted to the computational analysis of sRNA-seq data, in coordination with the bioinformatics platform of Génopole Toulouse-Midi-Pyrénées.

7.1.2. ADT

- 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.

7.2. European Initiatives

7.2.1. 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.

7.3. International Research Visitors

7.3.1. Visits of International Scientists

7.3.1.1. Internships

- Inria MITACS 3-month internship of D. Martchenko (PhD student, Trent University)

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events Organisation

8.1.1.1. General Chair, Scientific Chair

- H. Touzet was the general chair for the satellites of RECOMB 2018 (Paris, April 2018)

8.1.2. Scientific Events Selection

8.1.2.1. Chair of Conference Program Committees

- RECOMB-Seq 2018 (R. Chikhi)

8.1.2.2. Member of the Conference Program Committees

- RECOMB-CG 2018 (J.-S. Varré)
- WABI 2018 (H. Touzet, M. Salson)
- ACM-BCB 2018 (R. Chikhi)
- HiCOMB 2018 (R. Chikhi)

8.1.2.3. Reviewer

- WABI 2018 (L. Noé, J.-S. Varré, R. Chikhi)
- ECCB 2018 (R. Chikhi)

8.1.3. Journal

8.1.3.1. Member of the Editorial Boards

- Review Editor in Frontiers in Genetics - Bioinformatics and Computational Biology (J.-S. Varré)

8.1.3.2. Reviewer - Reviewing Activities

- Algorithms (L. Noé)
- Bioinformatics (L. Noé, M. Salson, R. Chikhi)
- Frontiers in Immunology (M. Salson)
- Nucleic Acids Research (R. Chikhi)
- GigaScience (R. Chikhi)
- ACM Computing Surveys (R. Chikhi)
- Genome Biology (R. Chikhi)
- Journal of Discrete Algorithms (R. Chikhi)

8.1.4. Scientific Expertise

- Scientific consulting for Clarity Genomics start-up – Belgium (R. Chikhi)

8.1.5. Research Administration

- 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 core facility, bilille (H. Touzet)
- Member of the CRIStAL Laboratory council (H. Touzet)

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Teaching in computer science:

License: J.-S. Varré, *Programming and algorithms*, 36h, L2 Computer Science, Univ. Lille.

License: J.-S. Varré, *Object oriented programming*, 36h, L2 Computer Science, Univ. Lille.

License: J.-S. Varré, *Algorithms and data structures*, 50h, L2 Computer science, Univ. Lille.

License: J.-S. Varré, *System*, 84h, L3 Computer science, Univ. Lille.

License: P. Marijon *Databases*, 36h, L3 Computer science, Univ. Lille.

License: Q. Bonenfant *Databases*, 36h, L3 Computer science, Univ. Lille.

Licence: M. Pupin, *Programming (Python)*, 78h, L1 Sciences, Univ. Lille.

Licence: M. Pupin, *occupational integration*, 30h, L3 computer science, Univ. Lille.

Master: : M. Pupin, *Programming (JAVA)*, 24h, M1 “Mathématiques et finance”, Univ. Lille.

License: M. Salson, *Programming (Python)*, 42h, L1 Sciences, Univ. Lille.

License: M. Salson, *Coding and information theory*, 63h, L2 Computer science, Univ. Lille.

Master: M. Salson, *Software project*, 40h, M1 Computer science, Univ. Lille.

Teaching in bioinformatics:

- Master: M. Pupin, M. Salson *Bioinformatics*, 34h, M1 ”Biologie-Santé“, Univ. Lille.
- Master: M. Salson, *Algorithms for life sciences*, 20h, M2 Complex models, algorithms and data, Univ. Lille.
- Master: R. Chikhi, *Bioinformatics*, 20h, M1 Computer Science, Univ. Lille.

Teaching in skeptical thinking:

- Master: M. Salson, *Skeptical thinking*, 27h, M2 Journalist and Scientist, ESJ, Univ. Lille.

Formation for academics:

- Bilille permanent training: C. Saad (*Variants*, 13h), R. Chikhi (*De novo assembly* and *Metagenomics de novo assembly*, 8h), H. Touzet (*DNA analyis* and *Metagenomics*, 8h), M. Salson (*RNA-seq analysis*, 1h).
- (JC)2BIM summer school: H. Touzet organized a national summer school in bioinformatics (5-8 June, Frejus), that gathered 30 participants and 12 trainers (among them, R. Chikhi and M. Salson): http://www.gdr-bim.cnrs.fr/?page_id=560.
- Workshop on Genomics: R. Chikhi (*de novo assembly & k-mers*, 8h), Czech Republic, 2 weeks, 80 participants: http://evomics.org/workshops/workshop-on-genomics/2018-workshop-on-genomics-cesky-krumlov/.
- CGSI summer school: R. Chikhi (*de novo assembly, metagenomics*, two keynote lectures), Los Angeles, 4 weeks, 100 participants: http://computationalgenomics.bioinformatics.ucla.edu.

8.2.2. Teaching administration

Head of the licence semester “Computer Science – S3 Harmonisation (S3H)”, Univ. Lille (L. Noé).

Member of faculty council (M. Pupin, J.-S. Varré).

Head of the 3rd year of licence of computer science, Univ. Lille (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 (M. Pupin).

Head of the *Informatique au féminin*, Univ. Lille (M. Pupin).

8.2.3. Supervision

PhD: T. Rocher, Indexing VDJ recombinations in lymphocytes for leukemia follow-up, February 2018, M. Giraud, M. Salson.

PhD: C. Saad, Caractérisation des erreurs de séquençage non aléatoires, application aux mosaïques et tumeurs hétérogènes, September 2018, M.-P. Buisine, H. Touzet, J. Leclerc, L. Noé, M. Figeac.

PhD in progress: P. Marijon, Analyse de graphes d'assemblage issus du séquençage ADN troisième génération, 2016, R. Chikhi, J-S. Varré.

PhD in progress: Q. Bonenfant, Algorithmes pour l'analyse de séquençage ARN troisième génération, 2017/11/15, L. Noé, H. Touzet.

8.2.4. Juries

- H. Touzet was member of the PhD juries of Camille Marchet (University of Rennes 1), Magali Dancette (University of Lyon 1), Aurélien Quillet (University of Rouen).
- H. Touzet was member of hiring committees (professors) at University of Lille and University of Caen.
- J.-S. Varré was member of a hiring committee (professors) at University of Lille.
- L. Noé was member of a hiring committee (associate professor) at University of Lille.
- R. Chikhi was member of the PhD juries of Florian Plaza Onate (INRA MetaGenoPolis), Sorina Maciuca (Wellcome Trust, UK).
- R. Chikhi was member of a hiring committee (research engineer) at Institut Pasteur, Paris.

8.3. Popularization

8.3.1. Internal or external Inria responsibilities

- Member of the CDT for Inria Lille (M. Pupin).
- Member of an Inria hiring committee (young researcher), Inria LNE (M. Pupin)

8.3.2. Education

- M. Pupin is the new leader of the collective *Informatique au féminin* from University of Lille, which was launched four years ago and whose goal is to organise computer science initiatives that reach teenage girls and female students. Among other actions, she is 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. 50 teenage girls were supervised by 11 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.
- 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", visit of high school students in the team (M. Salson).

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] T. ROCHER. *Compressing and indexing labeled sequences*, Université de Lille, February 2018, https://tel.archives-ouvertes.fr/tel-01758361

[2] C. SAAD. *Characterization of non-random sequencing errors, application to mosaicism and heterogeneous tumors*, Université de Lille Nord de France, September 2018, https://tel.archives-ouvertes.fr/tel-01936291

Articles in International Peer-Reviewed Journals

[3] V. G. CRAWFORD, A. KUHNLE, C. BOUCHER, R. CHIKHI, T. GAGIE. *Practical dynamic de Bruijn graphs*, in "Bioinformatics", June 2018 [*DOI :* 10.1093/BIOINFORMATICS/BTY500], https://hal.archives-ouvertes.fr/hal-01935559

[4] T. MARSCHALL, M. MARZ, T. ABEEL, L. DIJKSTRA, B. E. DUTILH, A. GHAFFAARI, P. KERSEY, W. P. KLOOSTERMAN, V. MAKINEN, A. M. NOVAK, B. PATEN, D. PORUBSKY, E. RIVALS, C. ALKAN, J. A. BAAIJENS, P. I. W. DE BAKKER, V. BOEVA, R. J. P. BONNAL, F. CHIAROMONTE, R. CHIKHI, F. D. CICCARELLI, R. CIJVAT, E. DATEMA, C. M. V. DUIJN, E. E. EICHLER, C. ERNST, E. ESKIN, E. GARRISON, M. EL-KEBIR, G. W. KLAU, J. O. KORBEL, E.-W. LAMEIJER, B. LANGMEAD, M. MARTIN, P. MEDVEDEV, J. C. MU, P. NEERINCX, K. OUWENS, P. PETERLONGO, N. PISANTI, S. RAHMANN, B. RAPHAEL, K. REINERT, D. DE RIDDER, J. DE RIDDER, M. SCHLESNER, O. SCHULZ-TRIEGLAFF, A. D. SANDERS, S. SHEIKHIZADEH, C. SHNEIDER, S. SMIT, D. VALENZUELA, J. WANG, L. WESSELS, Y. ZHANG, V. GURYEV, F. VANDIN, K. YE, A. SCHÖNHUTH. *Computational pan-genomics: status, promises and challenges*, in "Briefings in Bioinformatics", 2018, vol. 19, n^{O} 1, pp. 118-135 [*DOI :* 10.1093/BIB/BBW089], https://hal.inria.fr/hal-01390478

[5] M. PUPIN, A. FLISSI, P. JACQUES, V. LECLÈRE. *Bioinformatics tools for the discovery of new lipopeptides with biocontrol applications*, in "European Journal of Plant Pathology", July 2018 [*DOI :* 10.1007/s10658-018-1544-2], https://hal.archives-ouvertes.fr/hal-01937890

[6] T. ROCHER, M. GIRAUD, M. SALSON. *Indexing labeled sequences*, in "PeerJ Computer Science", 2018, vol. 4, pp. 1-14 [*DOI :* 10.7717/PEERJ-CS.148], https://hal.archives-ouvertes.fr/hal-01743104

[7] C. SAAD, L. NOÉ, H. RICHARD, J. LECLERC, M.-P. BUISINE, H. TOUZET, M. FIGEAC. *DiNAMO: highly sensitive DNA motif discovery in high-throughput sequencing data*, in "BMC Bioinformatics", December 2018, vol. 19, n^{O} 1 [*DOI :* 10.1186/s12859-018-2215-1], https://hal.inria.fr/hal-01881466

Conferences without Proceedings

[8] R. CHIKHI, V. JOVIČIĆ, S. KRATSCH, P. MEDVEDEV, M. MILANIC, S. RASKHODNIKOVA, N. VARMA. *Bipartite Graphs of Small Readability*, in "COCOON 2018 - The 24th International Computing and Combinatorics Conference", Qingdao, China, July 2018, https://arxiv.org/abs/1805.04765 , https://hal.archives-ouvertes.fr/hal-01935562

[9] A. KUOSMANEN, T. PAAVILAINEN, T. GAGIE, R. CHIKHI, A. I. TOMESCU, V. MAKINEN. *Using Minimum Path Cover to Boost Dynamic Programming on DAGs: Co-Linear Chaining Extended*, in "RECOMB 2018 - 22nd Annual International Conference on Research in Computational Molecular Biology", Paris, France, April 2018, https://arxiv.org/abs/1705.08754 , https://hal.archives-ouvertes.fr/hal-01935568

[10] P. MARQUET, M. PUPIN, Y. SECQ. *L codent, L créent: créations numériques artistiques pour démystifier l'informatique... au féminin! (descriptif d'atelier)*, in "Didapro 7 – DidaSTIC. De 0 à 1 ou l'heure de l'informatique à l'école", Lausanne, Switzerland, February 2018, pp. 1-2, https://hal.archives-ouvertes.fr/hal-01753402

Scientific Popularization

[11] R. DAVID, L. MABILE, M. YAHIA, A. CAMBON-THOMSEN, A.-S. ARCHAMBEAU, L. BEZUIDENHOUT, S. BEKAERT, G. BERTIER, E. BRAVO, J. CARPENTER, A. COHEN-NABEIRO, A. DELAVAUD, M. DE ROSA, L. DOLLÉ, F. GRATTAROLA, F. MURPHY, S. PAMERLON, A. SPECHT, A.-M. TASSÉ, M. THOMSEN, M. ZILIOLI. *Operationalizing and evaluating the FAIRness concept for a good quality of data sharing in Research: the RDA-SHARC-IG (SHAring Rewards and Credit Interest Group)*, November 2018, assemblée MaDICS 2018, Poster [*DOI :* 10.5281/ZENODO.1745374], https://hal.archives-ouvertes.fr/hal-01929834

Other Publications

[12] R. CHIKHI, A. SCHÖNHUTH. *Dualities in Tree Representations*, November 2018, https://arxiv.org/abs/1804.04263 - CPM 2018, extended version, https://hal.archives-ouvertes.fr/hal-01935566