IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2016

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:

6.2.7. - High performance computing
7.2. - Discrete mathematics, combinatorics
7.9. - Graph theory

Other Research Topics and Application Domains:

1.1.6. - Genomics
1.1.7. - Immunology
1.1.8. - Evolutionnary biology
1.1.9. - Bioinformatics
1.1.13. - Plant Biology
1.1.14. - Microbiology
1.2. - Ecology
1.2.1. - Biodiversity
2.2.3. - Cancer

1. Members

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. Lille I, Associate Professor]
Valérie Leclère [Univ. Lille I, Associate Professor, until Aug 2016, HDR]
Laurent Noé [Univ. Lille I, Associate Professor]
Maude Pupin [Univ. Lille I, Associate Professor, HDR]
Mikaël Salson [Univ. Lille I, Associate Professor]
Jean-Stéphane Varré [Univ. Lille I, Professor, HDR]

Engineers

Areski Flissi [CNRS]
Isabelle Guigon [CNRS, until Mar 2016]
Ryan Herbert [Inria]
Aurélien Béliard [CHU Lille, from Dec 2016]
Juraj Michalik [CNRS, until Mar 2016]

PhD Students

Yoann Dufresne [Univ. Lille I]
Pierre Marijon [Inria, from oct 2016]
Pierre Pericard [Univ. Lille I]
Tatiana Rocher [Univ. Lille I]
Chadi Saad [Univ. Lille II]

Léa Siegwald [CIFRE Gènes Diffusion]
Christophe Vroland [CNRS, until May 2016]

Post-Doctoral Fellow
Benjamin Momège [Inria]

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

The software SortMeRNA, developed by the team, has reached the number of 100 labs worldwide that have been using it to analyze their sequencing data. SortMeRNA is able to deal with large metagenomics projects with multiple applications in health (gut microbiome,...), environment (sea, lakes, soil,...), biotechnologies (bio-films,...). The first version was released at the end of 2012, and it is still under active maintenance.

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: Rayan Chikhi, Antoine Limasset and Paul Medvedev
- Contact: Rayan Chikhi
- URL: https://github.com/GATB/bcalm

6.2. NORINE

Nonribosomal peptides resource
KEYWORDS: Bioinformatics - Biotechnology - Biology - Genomics - Graph algorithmics - Chemistry - Knowledge database - Drug development - 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. Norine is freely accessible from the following URL: http://bioinfo.lifl.fr/NRP.

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 easly submit new nonribosomal peptides, Smiles2monomers (s2m), a tool that deciphers the monomeric structure of polymers from their chemical structure.

- Participants: Maude Pupin, Areski Flissi, Valerie Leclere, Laurent Noe, Yoann Dufresne, Juraj Michalik and Stéphane Janot
- Partners: CNRS - Institut Charles Viollette - Université Lille 1 v
- Contact: Maude Pupin
- URL: http://bioinfo.lille.inria.fr/NRP

6.3. Olo

KEYWORDS: Bioinformatics - Indexation - Sequence alignment - Biological sequences - Approximate string matching

SCIENTIFIC DESCRIPTION

Approximate string matching of short sequences in a text often starts by a filtering step. That step relies on seed searching, which are shorter than the pattern. Usually in those seeds the number of errors is constrained, to allow more efficient computations. We designed the 01*0 seeds which offer a good trade-off between the number of false positives and filtering time.

FUNCTIONAL DESCRIPTION

We applied the 01*0 seeds to the similarity search of miRNA targets in a reference genome (Bwolo software) and to the similarity search between a pre-miRNA and mature miRNAs (Piccolo software).

- Participants: Sébastien Bini, Mikaël Salson, Hélène Touzet and Christophe Vroland
- Partners: CNRS - Université Lille 1
- Contact: Helene Touzet
- URL: http://bioinfo.lifl.fr/olo/

6.4. Vidjil

High-Throughput Analysis of V(D)J Immune Repertoire

KEYWORDS: Bioinformatics - NGS - Indexation - Cancer - Drug development

SCIENTIFIC DESCRIPTION

Vidjil is made of three components: an algorithm, a visualisation 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 gather 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: Mathieu Giraud, Mikaël Salson, Marc Duez, Ryan Herbert, Tatiana Rocher and Florian Thonier
- Partners: CHRU Lille - CNRS - Inria - Université de Lille
- Contact: Mathieu Giraud
- URL: http://www.vidjil.org

7. New Results

7.1. Approximate pattern matching

The problem of measuring the similarity between two strings arises in many areas of sequence analysis. A common metric for it is the *Levenshtein distance*. This distance is defined as the smallest number of substitutions, insertions, and deletions of symbols required to transform one of the words into the other. We have investigated the basic problem of the size of the neighborhood of a given pattern P: count how many strings are within a bounded distance of a fixed reference string. There has been no efficient algorithm for calculating it so far. We have proposed a dynamic programming algorithm that scales linearly with the size of the pattern P. For that, we have introduced a new variant of the universal Levenshtein automaton, that is interesting by itself and that can have many other applications in text algorithms [31].

We have also addressed the related problem of approximate pattern matching: Given a text T and a pattern P, find all locations in T that differ by at most k errors (in the sense of the Levenshtein distance) from P. We have proposed a new kind of seeds (the 01^*0 seeds) that combines exact parts and parts with a fixed number of errors, and that are specifically well-suited for short DNA motifs with high error-rate. We have demonstrated the applicability of those seeds on two main case studies : pattern matching on a genomic scale with a Burrows-Wheeler transform, and multi-pattern matching with indexation of the set of patterns [30].

7.2. Parallel algorithm for de Bruijn graph compaction

Constructing a *de Bruijn graph* is an important step in the analysis of NGS data. This data structure is used in several applications, such as *de novo* assembly, variant detection, and transcriptome quantification. However, the representation of this graph often consumes prohibitive amounts of memory for large datasets. An operation, called compaction, enables to represent the graph more efficiently. However, so far, there was no algorithm for compacting the graph quickly and in low memory.

Along with colleagues at Inria Rennes and at Penn State University, we introduced a parallel algorithm and an implementation, BCALM 2, for constructing directly a compacted de Bruijn graph given a set of reads. Our results show that this algorithm enables to construct the graph for very large datasets, such as the spruce and pine genomes, in reasonable time and memory on a single machine. This represents a performance improvement of two orders of magnitude compared to previously available methods. BCALM 2 is open-source and was published at ISMB 2016 [20].

7.3. Range minimum query

The *range minimum query* problem consists in finding the minimum value inside any queried range of a preprocessed integer sequence. Several methods exist to compute the minimum in constant time, using almost the theoretical minimal amount of space. Those methods consist in splitting the problem in several subproblems and precomputing the solutions for them.

With Alice Héliou (AMIB Inria team, Saclay), Martine Léonard and Laurent Mouchard (LITIS, Rouen), we designed a new method, which is worse in terms of time complexity [24]. Our solution relies on a totally different concept as previous ones: We only store the values that are local minima. This approach is therefore simple and can, on specific inputs, require much less memory than the general theoretical minimal bound. Moreover the simplicity of the method can be easily adapted to allow updates in the original integer sequence.

7.4. Coding isoform structures

Our researches on gene isoform structures started in 2014 with the CG-Alcode Associated Team and in collaboration with Anne Bergeron from the LACIM (Montréal, Canada). We aimed at defining better definitions of isoform orthology at the coding level, which are based on the preservation of all the exon junctions in two orthologous isoforms. This estimation is achieved at the gene level, where sequence homology is detected for both exons and their flanking intronic splice sites [19]. The approach largely outperforms competing programs in terms of precision and recall. Using the successive releases of the ten years old CCDS database, we show that the discovery rate of orthologous isoforms between human and mouse is growing continuously and that it displays no sign of completion.

7.5. Nonribosomal peptides

We were invited to contribute in a volume of "Methods in Molecular Biology" by authoring a chapter focusing on NRPS biosynthesis. This chapter [32] was about the use of the Norine platform (developed by the team) and other bioinformatics tools for the analysis of nonribosomal peptide synthetases and their products. We invited our collaborator from Denmark, Tilmann Weber, to complete this chapter with the introduction of his tool, antiSMASH.

We annotated 48 genomes of *Burkholderia* species using our annotation protocol, that starts from a genome sequence and goes to the predicted nonribosomal peptides. We have predicted 161 gene clusters producing nonribosomal peptides, leading to the synthesis of not only already known peptides, but also new ones [22] with potential applications in biocontrol.

A new version of the Norine interface is now available. The form to query the annotations is now flexible and dynamic. The user can build his own query to search for annotations in several fields combined by boolean operators. Moreover, the database structure has been modified to allow, among others, a hierarchical representation of the NRPS taxonomy. Finally, the MyNorine tool has been enhanced and updated to take into account these changes and the description page of the peptides has been reorganized.

7.6. 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 [21].

In 2016, with our colleagues at Lille hospital, we published two articles in haematological journals to detail our method for the diagnosis [23] and for the follow-up [28] of the acute lymphoblastic leukemia using high-throughput sequencing. Our results also show what those new techniques, together with bioinformatics software, bring in a routine practice. Being a full platform with metadata storage, Vidjil is used on a regular basis by about 20 laboratories around the world. In France, the majority of diagnosis samples from acute lymphoblastic leukemia patients are now analyzed using Vidjil.

7.7. Assembly of the giraffe genome and the gorilla Y-chromosome

We collaborated with two labs from the Pennsylvania State Institute (Cavener Lab and Makova Lab) for practical analysis of DNA sequencing data. The first collaboration led to the publication of the giraffe genome in Nature Communication [18]. In this article our contribution was to provide the first draft-quality whole-genome sequences of the giraffe and the okapi. The second collaboration was about assembling the Y-chromosome of the gorilla using a novel sequencing strategy as well as novel computational tools. This work was published in Genome Research [29].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

The PhD thesis of Léa Siegwald is funded by a CIFRE contract with the biotechnology company Gènes Diffusion.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. 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.

9.1.2. ADT

- ADT Vidjil (2015–2017): The purpose of this ADT is to strengthen Vidjil development and to ensure a better diffusion of the software by easing its installation, administration and usability. This will enable the software to be well suited for a daily clinical use. The software is already used in test on our own web server (more than 5,000 samples processed by now). Vidjil is now used in a routine practice by three French hospitals and one German hospital. By the end of the ADT, we expect this number to increase and the software to be directly installed inside some hospitals.

9.2. European Initiatives

9.2.1. Collaborations in European Programs, Except FP7 & H2020

International ANR RNAlands (2014-2017): 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.

9.3. International Initiatives

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

9.3.1.1. CG-ALCODE

Title: Comparative Genomics for the analysis of gene structure evolution: ALternative CODing in Eukaryote genes through alternative splicing, transcription, and translation.

International Partner (Institution - Laboratory - Researcher):

Université du Québec À Montréal (Canada) - Laboratoire de combinatoire et d'informatique mathématique (LaCIM) - Anne Bergeron

From 2014 to 2017

The aim of this Associated Team is the development of comparative genomics models and methods for the analysis of eukaryote genes structure evolution. Our goal is to answer very important questions arising from recent discoveries on the major role played by alternative transcription, splicing, and translation, in the functional diversification of eukaryote genes. Two working meeting took place in 2016. S. Blanquart and J.-S. Varré met A. Bergeron and K. Swenson in Montpellier, from 13th to 15th of April. J.-S. Varré and K. Swenson met A. Bergeron in Montréal, from 8th to 19th of November.

9.3.2. Inria International Partners

9.3.2.1. Informal International Partners

- *Astrid Lindgrens Hospital, Stockholm University:* Collaboration with Anna Nilsson and Shanie Saghafian-Hedengren on RNA sequencing of stromal cells.
- *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 and Jasmijn Baaijens on succinct data structures and algorithms for the assembly of viral quasispecies.
- *Department of Statitics, North Carolina State University:* Collaboration with Donald E. K. Martin on spaced seeds coverage.
- *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.
- *Institut für Biophysik und physikalische Biochemie, University of Regensburg:* Collaboration with Rainer Merkl on ancestral sequence inference and synthesis.
- *Institute of Biosciences and Bioresources, Bari:* Collaboration with Nunzia Scotti on the assembly of plant mitochondrial genomes.
- *Makova lab, The Pennsylvania State University:* Collaboration with Kateryna Makova and Samarth Rangavittal on the assembly of the gorilla Y chromosome, and visualisation of assembly graphs.
- *Medvedev lab, The Pennsylvania State University:* Collaboration with Paul Medvedev on algorithms for constructing de Bruijn graphs.

- *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 Marc Duez, 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.

9.3.3. 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).

9.4. International Research Visitors

9.4.1. Visits of International Scientists

- Dr. Alexander Schoenhuth (group leader, *CWI Amsterdam*) and Jasmijn Baaijens (PhD student, *CWI Amsterdam*).

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organization

10.1.1.1. General Chair, Scientific Chair

- RepSeq 2016, workshop on immune repertoire sequencing at ECCB 2016 (M. Giraud, M. Salson).

10.1.1.2. Member of the Organizing Committees

- SMPGD, Statistical Methods in Post Genomic Data (H. Touzet).

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- WABI 2016 (H. Touzet).
- RECOM-seq 2016 (H. Touzet).

10.1.2.2. Reviewer

- ECCB 2016 (J.-S. Varré, R. Chikhi).
- PSC 2016 (M. Salson).
- RECOMB-CG 2016 (J.-S. Varré).
- TCBB 2016 (J.-S. Varré).
- RECOMB 2016 (R. Chikhi).

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing Activities

- Bioinformatics (M. Salson, J.-S. Varré, R. Chikhi, H. Touzet).
- PLoS Genetics (R. Chikhi).
- Genome Research (R. Chikhi).

- Nucleic Acids Research (R. Chikhi).
- Journal of Discrete Algorithms (M. Salson).
- Briefing in Bioinformatics (H. Touzet).

10.1.4. Invited Talks

- RegPep2016 (regulated peptides) in Rouen (France), Symposium 8 "In silico approaches to peptide identification and design" (M. Pupin).
- SMPGD keynote in Lille (France), (R. Chikhi).
- National meeting of GDR Informatique Mathématique in Paris (H. Touzet).
- MatBio 2016 in London (H. Touzet).
- Summer school in metagenomics in Paris (H. Touzet).

10.1.5. Scientific Expertise

- Reviewer for a Swiss National Science Foundation grant (R. Chikhi).
- Member of the scientific committee of the national program Environmics (H. Touzet).
- Reviewer for a CRSNG grant (H. Touzet).
- Reviewer for labex CIMI PhD grant (H. Touzet).
- Member of three HCERES committees: L2N (LINA and IRCCyN, Nantes), Loria (Nancy), I2M (Marseille) (H. Touzet).

10.1.6. Research Administration

- Member of the CUB for Inria Lille (S. Blanquart).
- Member of the Charles Viollette Institute Laboratory council (V. Leclère).
- Member of the Charles Viollette Institute scientific committee (V. Leclère).
- Member of the scientific operational committee of Xperium, Univ. Lille 1 (V. Leclère).
- Member of the Inria local committee for technology development (M. Pupin).
- Member of the executive council of the IFB, Institut Français de Bioinformatique, (M. Pupin).
- Member of the Inria local committee for the IT users (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).
- Vice-head of the Lille Bioinformatics platform, bilille (H. Touzet).
- Member of the CRIStAL Laboratory council (H. Touzet).
- Member of the CRIStAL scientific council, coordinator of the thematic group “Modeling for life sciences” (J.-S. Varré).

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Teaching in computer science:

- Master: Y. Dufresne, *Algorithmics and complexity*, 36h, M1 Computer Science, Univ. Lille 1.
- License: Y. Dufresne, *Oriented object design*, 42h, 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.
- 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: P. Pericard, *Data structures*, 18h, L3 Polytech'Lille, Univ. Lille 1.
- License: P. Pericard, *Introduction to programming (C)*, 34h, L3 Polytech'Lille, Univ. Lille 1.
- License: P. Pericard, *Programming (C)*, 22h, L3 Polytech'Lille, Univ. Lille 1.
- License: P. Pericard, *Databases*, 22h, M1 Polytech'Lille, Univ. Lille 1.
- License: M. Pupin, *Introduction to programming (Python)*, 78h, L1 Computer science, Univ. Lille 1.
- License: M. Pupin, *Professional project*, 18h, L3 Computer science, Univ. Lille 1.
- Master: M. Pupin, *Introduction to programming (JAVA)*, 24h, M1 Mathématiques et finance, Univ. Lille 1.
- License: T. Rocher, *Algorithmics and programming*, 32h, L3 Polytech'Lille, Univ. Lille 1.
- License: T. Rocher, *Algorithmics and programming (remedial course)*, 8h, L3 Polytech'Lille, Univ. Lille 1.
- License: T. Rocher, *Databases*, 26h, L3 Polytech'Lille, 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.
- Master: M. Salson, *Skeptical thinking*, 18h, M2 Journalist and Scientist, ESJ, Univ. Lille 1.
- License: M. Salson, *Coding and information theory*, 63h, L2 Computer science, Univ. Lille 1.
- License: J.-S. Varré, *Web programming*, 36h, L2 Computer Science, Univ. Lille 1.
- License: J.-S. Varré, Y. Dufresne, *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*, 36h, L3 Computer science, Univ. Lille 1.
- Master: J.-S. Varré, Y. Dufresne, *Software project*, 24h, M1 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.
- License: V. Leclère, *Biotechnology*, 24h, L3 Biology, Univ. Lille 1.
- Master: L. Noé, *Bioinformatics*, 40h, M1 Biotechnologies, Univ. Lille 1.
- Master: M. Pupin *Bioinformatics*, 40h, M1 Biology and Biotechnologies, Univ. Lille 1.
- Master: M. Salson, *Algorithms for life sciences*, 18h, M2 Complex models, algorithms and data, Univ. Lille 1.

Teaching in biology:

- Master, V. Leclère, *Mycology, secondary metabolites, food microbiology*, 37 h, M1 Biology, Univ. Lille 1.

10.2.2. Teaching administration

- Head of the computer science modules in the 1st year of licence, univ. Lille 1 (M. Pupin).
- Head of the licence semester "Computer Science – S3 Harmonisation (S3H)", univ. Lille 1 (L. Noé).
- Member of UFR IEEA 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).
- Member of UFR Biologie council (V. Leclère).
- Head of the master "Innovations en biotechnologies végétales, enzymatiques et microbiennes", univ. Lille 1 (V. Leclère).

10.2.3. Supervision

- PhD : C. Vroland, Algorithmique pour la recherche de motifs approchée et application à la recherche de cibles de microARN, univ. Lille 1, 2016/05/18, H. Touzet, V. Castric, M. Salson.
- PhD : Y. Dufresne, Algorithmes pour l'annotation automatique de peptides non-ribosomiques, univ. Lille 1, 2016/12/01, M. Pupin, L. Noé.
- PhD in progress: P. Pericard, Methods for taxonomic assignation in metagenomics, 2013/11/01, 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: L. Siegwald, Bionformatic analysis of Ion Torrent metagenomic data, 2014/01/03, H. Touzet, Y. Lemoine (Institut Pasteur de Lille).
- PhD in progress: P. Marijon, Graph assembly analysis for third generation sequencing data, 2016/10/01, J.-S. Varré, R. Chikhi (Institut Pasteur de Lille).

10.2.4. Juries

- Member of the HDR committee of Laurent Mouchard (Univ. Rouen, J.-S. Varré).
- Member of the PhD committee of Qassin Esmael (Univ. Lille 1, M. Pupin, V. Leclère).
- Member of the PhD committee of Souhir Sabri (Univ Montpellier, V. Leclère).
- Member of the PhD committee of Wahiba Chaara (Univ. Paris 6, M. Giraud).
- Member of the PhD committee of Leandro Ishi (Univ. Lyon, R. Chikhi).
- Member of the PhD committee of Jerome Audoux (Univ. Montpellier, R. Chikhi, M. Salson).
- Member of the PhD jury of Clara Benoit (Univ. Lyon, R. Chikhi).
- Member of the HDR jury of Morgane Thomas-Chollier (IBENS, Ecole Normale Supérieure Paris, H. Touzet).
- Member of the HDR jury of Annie Chateau (LIRMM, Université de Montpellier, H. Touzet).
- Member of the HDR jury of Pierre Peterlongo (Inria Rennes, H. Touzet).
- Member of the HDR jury of Séverine Bérard (ISEM, Université de Montpellier, H. Touzet).
- Member of the PhD jury of Thomas Hume (LaBRI, Université Bordeaux 1, H. Touzet).
- Member of the PhD jury of Aymeric Antoine-Lorquin (IRISA, Université de Rennes 1, H. Touzet).
- Member of the hiring committee MdC of Univ. Nancy (M. Pupin).
- Member of the hiring committee MdC of Univ Lille 1 (V. Leclère).
- Member of the hiring committee professor of Univ. Rouen (H. Touzet).

- Member of the hiring committee Research Engineer of Univ. Paris-Diderot (M. Pupin).
- Member of the hiring committee of research engineer of Univ Lille 1 (V. Leclère).

10.3. Popularization

The team has always been very active in popularizing computational biology and computer science in general.

- M. Salson participated in an exchange with scientific journalists organized by the French association of scientific journalists (AJSPI). A journalist was hosted during one week in the team and M. Salson spent one week in the newsroom of *La Recherche*, a French science magazine.
- Within a project on skeptical thinking with a popularization association “Les Petits Débrouillards”, M. Salson is part of the monitoring committee and gave lectures to social workers.
- 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”, booth about computer science unplugged for high school girls, booth at Xperium about development of biopesticides (including a demo on the use of Norine), plenary presentations at the “Day for Programming and Algorithmic Teaching”, presentations at “Salon de l’étudiant”, visit of high school students in the team (M. Giraud, M. Pupin, M. Salson, J.-S. Varré, R. Chikhi, V. Leclère)

11. 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° 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° 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° 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° 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° 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° 3, pp. 190-204, https://hal.archives-ouvertes.fr/hal-01228130

Publications of the year

Articles in International Peer-Reviewed Journals

[18] M. AGABA, E. ISHENGOMA, W. C. MILLER, B. C. MCGRATH, C. N. HUDSON, O. C. BEDOYA REINA, A. RATAN, R. BURHANS, R. CHIKHI, P. MEDVEDEV, C. A. PRAUL, L. CAVENER, B. WOOD, H. ROBERTSON, L. PENFOLD, D. R. CAVENER. *Giraffe genome sequence reveals clues to its unique morphology and physiology*, in "Nature Communications", May 2016, vol. 7 [*DOI :* 10.1038/NCOMMS11519], https://hal.archives-ouvertes.fr/hal-01395703

[19] S. BLANQUART, J.-S. VARRÉ, P. GUERTIN, A. PERRIN, A. BERGERON, K. M. SWENSON. *Assisted transcriptome reconstruction and splicing orthology*, in "BMC Genomics", 2016, vol. 17, n° 786 [*DOI :* 10.1186/S12864-016-3103-6], https://hal.inria.fr/hal-01396410

[20] 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° 12, pp. i201 - i208 [*DOI :* 10.1093/BIOINFORMATICS/BTW279], https://hal.archives-ouvertes.fr/hal-01395704

[21] M. DUEZ, M. GIRAUD, R. HERBERT, T. ROCHER, M. SALSON, F. THONIER. *Vidjil: A Web Platform for Analysis of High-Throughput Repertoire Sequencing*, in "PLoS ONE", November 2016, vol. 11, n° 11 [*DOI :* 10.1371/JOURNAL.PONE.0166126], https://hal.archives-ouvertes.fr/hal-01397079

[22] Q. ESMAEEL, M. PUPIN, N. P. KIEU, G. CHATAIGNÉ, M. BÉCHET, J. DERAVEL, F. KRIER, M. HÖFTE, P. JACQUES, V. LECLÈRE. *Burkholderia genome mining for nonribosomal peptide synthetases reveals a great potential for novel siderophores and lipopeptides synthesis*, in "MicrobiologyOpen", May 2016, vol. 5, n° 3, pp. 512 - 526 [*DOI :* 10.1002/MBO3.347], https://hal.archives-ouvertes.fr/hal-01398944

[23] 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

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

[25] 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. D. 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. D. RIDDER, J. D. 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", October 2016 [*DOI :* 10.1093/BIB/BBW089], https://hal.inria.fr/hal-01390478

[26] 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° 1, pp. 231–248 [*DOI :* 10.1007/S10463-015-0540-Y], https://hal.inria.fr/hal-01237045

[27] K. SAHLIN, R. CHIKHI, L. ARVESTAD. *Assembly scaffolding with PE-contaminated mate-pair libraries*, in "Bioinformatics", March 2016, vol. 32, n° 13, pp. 1925 - 1932 [*DOI :* 10.1093/BIOINFORMATICS/BTW064], https://hal.archives-ouvertes.fr/hal-01396904

[28] 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

[29] M. TOMASZKIEWICZ, S. RANGAVITTAL, M. CECHOVA, R. C. SANCHEZ, H. W. FESCEMYER, R. HARRIS, D. YE, P. C. O'BRIEN, R. CHIKHI, O. A. RYDER, M. A. FERGUSON-SMITH, P. MEDVEDEV, K. D. MAKOVA. *A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y*, in "Genome Research", March 2016, vol. 26, n° 4, pp. 530 - 540 [*DOI :* 10.1101/GR.199448.115], https://hal.archives-ouvertes.fr/hal-01395702

[30] C. VROLAND, M. SALSON, S. BINI, H. TOUZET. *Approximate search of short patterns with high error rates using the 010 lossless seeds*, in "Journal of Discrete Algorithms", 2016, vol. 37, pp. 3-16 [*DOI :* 10.1016/J.JDA.2016.03.002], https://hal.archives-ouvertes.fr/hal-01360485

International Conferences with Proceedings

[31] H. TOUZET. *On the Levenshtein Automaton and the Size of the Neighborhood of a Word*, in "Language and Automata Theory and Applications", Prague, Czech Republic, A.-H. DEDIU, J. JANOUŠEK, C. MARTÍN-VIDE, B. TRUTHE (editors), Lecture Notes in Computer Sciences, Springer, 2016, vol. 9618, pp. 207-218 [*DOI :* 10.1007/978-3-319-30000-9_16], https://hal.archives-ouvertes.fr/hal-01360482

Scientific Books (or Scientific Book chapters)

[32] V. LECLÈRE, T. WEBER, P. JACQUES, M. PUPIN. *Bioinformatics Tools for the Discovery of New Nonribosomal Peptides*, in "Nonribosomal Peptide and Polyketide Biosynthesis", B. S. EVANS (editor), Methods in Molecular Biology, Humana Press, February 2016, vol. 1401, pp. 209-232 [*DOI :* 10.1007/978-1-4939-3375-4_14], https://hal.archives-ouvertes.fr/hal-01398960

Other Publications

[33] M. SALSON, A. CAILLAULT, M. DUEZ, Y. FERRET, A. FIEVET, M. KOTROVA, F. THONIER, P. VILLARESE, S. WAKEMAN, G. WRIGHT, M. GIRAUD. *A dataset of sequences with manually curated V(D)J designations*, 2016, Workshop Immune Repertoire Sequencing : Bioinformatics and Applications in Hematology and Immunology (RepSeq 2016), https://hal.archives-ouvertes.fr/hal-01331556