IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2015

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
1.2.1. - 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]

Faculty Members

Stéphane Janot [Univ. Lille I, Associate Professor]
Valérie Leclère [Univ. Lille I, Associate Professor, 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

Marc Duez [Univ. Lille II, until Jun 2015]
Areski Flissi [CNRS]
Isabelle Guigon [CNRS]
Ryan Herbert [Inria, from Oct 2015]
Alan Lahure [CNRS, until Dec 2015]
Juraj Michalik [CNRS]
Amandine Perrin [Inria, until Feb 2015]

PhD Students

Yoann Dufresne [Univ. Lille I]
Pierre Pericard [Univ. Lille I]
Tatiana Rocher [Univ. Lille I]
Chadi Saad [Univ. Lille II]
Léa Siegwald [CIFRE Genes Diffusion]
Christophe Vroland [Univ. Lille 1 until Oct 2015, CNRS from Oct 2015]

Post-Doctoral Fellow

Benjamin Momège [Inria, from Nov 2015]

Administrative Assistants

Natacha Oudoire [Inria]

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 modelled 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 modelled 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 labelled 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 modelled 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 analyse 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 analyse 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 lab 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

5.1.1. MyNorine invents the crowd sourcing for Non Ribosomal Peptides

For ten years, the team has been developing a unique knowledge base, Norine, dedicated to the modeling and analysis of Nonribosomal peptides (NRPs). NRPs are secondary metabolites produced by bacteria and fungi that represent a huge source of natural products with applications in agricultural or health areas. In January 2015, we have released a new version which contains several major advances. MyNorine is a user-friendly service, that allows to submit new NRPs and to edit existing ones [2]. It was tested and validated by a panel of expert users during an international workshop that we organized in Lille in October, and that attracted 32 attendees from 8 countries. Moreover, s2m is an innovative tools to infer the monomeric structure of the peptides [1].

5.1.2. 1,000 white blood cell samples processed by Vidjil

Vidjil is an open-source platform for the analysis of high-throughput sequencing data from lymphocytes developed by the team. In October 2014, we opened a web server to grant everyone an access to Vidjil, our white blood cell analysis software used for leukemia diagnosis and follow-up. For one year, Vidjil analyzed more than 1,000 samples totalling 5 billion DNA sequences. Our users come from about thirty hospitals and labs throughout the world [3]. About fifteen of them regularly submit new samples. In Lille, the hematology department of the hospital uses Vidjil to identify malignant white blood cells at diagnosis on every patient with acute leukemia.

6. New Software and Platforms

6.1. Iedera

Iedera : subset seed design tool
KEYWORDS: Computational biology - Sequence alignment - Spaced seeds
SCIENTIFIC DESCRIPTION

Iedera is a tool based on a weighted automata framework that enables to compute spaced seeds, associated probability distributions, scores/costs, counts, and even polynomials on a classical semi-ring framework. Seed design is computed by full enumeration or hill-climbing optimization.
FUNCTIONAL DESCRIPTION

Iedera is a tool to select and design subset seed and vectorized subset seed patterns. Spaced seeds and transition-constrained spaced seeds can be perfectly represented in the subset seed model.

- Participants: Laurent Noé, Grégory Kucherov and Mikhail Roytberg
- Partners: CNRS - Université de Nancy - Université de Lille
- Contact: Laurent Noé
- URL: http://bioinfo.lifl.fr/yass/iedera.php

6.2. NORINE

Nonribosomal peptides resource
KEYWORDS: Bioinformatics - Biotechnology - Biology - Genomics - Graph algorithmics - Chemistry - Knowledge database - Drug development - Computational biology
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 Leclère, Laurent Noé, Yoann Dufresne, Juraj Michalik and Stéphane Janot
- Partners: CNRS - Université Lille 1 - Institut Charles Viollette
- Contact: Maude Pupin
- URL: http://bioinfo.lille.inria.fr/NRP

6.3. ProCARs

PROgressive Contiguous Ancestral RegionS
KEYWORDS: Bioinformatics - Evolution - Metagenomics
SCIENTIFIC DESCRIPTION

ProCars is a program used to reconstruct ancestral gene orders as CARs (Contiguous Ancestral Regions) with a progressive homology-based method. The method runs from a phylogeny tree (without branch lengths needed) with a marked ancestor and a block file. This homology-based method is based on iteratively detecting and assembling ancestral adjacencies, while allowing some micro-rearrangements of synteny blocks at the extremities of the progressively assembled CARs. The method starts with a set of blocks as initial set of CARs, and detects iteratively the potential ancestral adjacencies between extremities of CARs, while building up the CARs progressively by adding, at each step, new non-conflicting adjacencies that induce the less homoplasy phenomenon. The species tree is used, in some additional internal steps, to compute a score for the remaining conflicting adjacencies, and to detect other reliable adjacencies, in order to reach completely assembled ancestral genomes.
FUNCTIONAL DESCRIPTION

ProCARs is a program used to reconstruct ancestral gene orders as CARs (Contiguous Ancestral Regions) with a progressive homology-based method. The method runs from a phylogeny tree with a marked ancestor and a block file.

- Participants: Aïda Ouangraoua, Samuel Blanquart, Jean-Stéphane Varré and Amandine Perrin
- Partners: CNRS - Université de Lille
- Contact: Jean-Stéphane Varré
- URL: http://bioinfo.lifl.fr/procars

6.4. SortMeRNA

KEYWORDS: Bioinformatics - NGS - Genomic sequence
SCIENTIFIC DESCRIPTION

SortMeRNA is a biological sequence analysis tool for metatranscriptomic and metagenomic data filtering, mapping and OTU-picking. The main application of SortMeRNA is filtering and mapping ribosomal RNA from NGS reads.

FUNCTIONAL DESCRIPTION The core algorithm is based on approximate seeds as well as an optimized text index data structure. It allows for fast and sensitive analyses of nucleotide sequences.

SortMeRNA takes as input a file of reads (fasta or fastq format) and one or multiple rRNA database file(s), and sorts apart rRNA and rejected reads into two files specified by the user. Optionally, it can provide high quality local alignments of rRNA reads against the rRNA database. SortMeRNA works with Illumina, 454, Ion Torrent and PacBio data, and can produce SAM and BLAST-like alignments. It is implemented in C++.

- Participants: Hélène Touzet, Laurent Noé and Evguenia Kopylova
- Contact: Hélène Touzet
- URL: http://bioinfo.lille.inria.fr/RNA/sortmerna/

6.5. 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 againt 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 a 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: CNRS - Inria - Université de Lille
- Contact: Mathieu Giraud
- URL: http://www.vidjil.org

6.6. Yass

KEYWORDS: Bioinformatics - Genomic sequence - Computational biology - Sequence alignment
SCIENTIFIC DESCRIPTION

As most of the heuristic DNA local alignment softwares (BLAST, FASTA, PATTERNHUNTER, BLASTZ, LAST...) YASS uses seeds to detect potential similarity regions, and then tries to extend them to actual alignments.

This genomic search tool uses multiple transition-constrained spaced seeds (most of the design of these seeds is provided by the Iedera tool) to search for more fuzzy repeats, such as non-coding DNA/RNA.

Main features of YASS are: (i) multiple, possibly overlapping seeds and a new hit criterion to ensure a good sensitivity/selectivity trade-off (ii) transition-constrained spaced seeds to improve sensitivity (transition mutations are purine to purine [AG] or pyrimidine to pyrimidine [CT]) (iii) using different scoring schemes with bit-score and E-value evaluated according to the sequence background frequencies (iv) parameterizable output filter for low complexity repeats (v) reporting of various alignment statistical parameters (mutation bias along triplets, transition/transversion), and (vi) post-processing step to group gapped alignments.

FUNCTIONAL DESCRIPTION

YASS is a genomic similarity search tool, for nucleic (DNA/RNA) sequences in fasta or plain text format : it produces local pairwise alignments.

- Participants: Laurent Noé and Grégory Kucherov
- Partners: CNRS - Université de Nancy - Université de Lille
- Contact: Laurent Noé
- URL: http://bioinfo.lifl.fr/yass

6.7. miRkwood

KEYWORDS: Bioinformatics - Genomics

SCIENTIFIC DESCRIPTION

miRkwood is a bioinformatic pipeline that allows for the fast and easy identification of microRNAs in plant genomes. It is both available as a webserver and a stand-alone software. It offers an user-friendly interface to navigate in the data, as well as many export options to allow the user to conduct further analyses on a local computer.

FUNCTIONAL DESCRIPTION

The method takes as input a set of small reads, that have been previously trimmed and aligned onto the reference genome. It identifies novel microRNAs on the basis of the distributions of reads and the potential of flanking genomic sequence to fold into a stem-loop secondary structure. Then the result is refined through a variety of additional complementary features that bring new evidence to the prediction: duplex stability, thermodynamic stability, phylogenetic conservation, repeats, etc.

- Participants: Hélène Touzet, Mohcen Benmounah, Jean-Frédéric Berthelot, Isabelle Guigon and Sylvain Legrand
- Contact: Hélène Touzet
- URL: http://bioinfo.lille.inria.fr/mirkwood/

7. New Results

7.1. Ancestral gene order reconstruction

In the field of **genomic rearrangement**, a topic of interest is to infer ancestral gene order from gene order known in extant species. The problem resumes to compute a set ancestral CARs (continuous ancestral regions) at a given node of a phylogeny. This work, initially published in a conference, was published this year in a journal [5].

7.2. Nonribosomal peptides

Norine is the unique and leading platform dedicated to computational biology analysis of nonribosomal peptides (NRPs). It is used by thousands of scientists all over the world to explore and better understand the diversity of the NRPs. To improve the data quality and quantity in Norine, we are now opening our resource to external contributors. To achieve this new challenge, we developed new tools (MyNorine, s2m) and communicate on our novelties.

- **Crowdsourcing.** To facilitate the submission of new nonribosomal peptides (NRPs) or modification of stored ones in Norine, we have developed a dedicated and user-friendly module named MyNorine [2]. It provides interactive forms to fill in the annotations with, for example, autocompletion and tools such as a monomeric structure editor. It has especially been designed for biologists and biochemists working on secondary metabolites to easily enrich the database with their own data.

- **Norine communication.** We advertise Norine by different promoting media. We organized an international workshop in Lille in October to teach biologists and biochemists how to annotate NRPs and their synthetases with bioinformatics tools such as Norine. It attracted 32 attendees from 8 countries. We participated, as invited contributors, to the special issue "Bioinformatics tools and approaches for synthetic biology" of the new journal Synthetic and Systems Biotechnology edited by KeAi Publishing, funded by Elsevier and Chinese Science Publishing & Media. Our article [6] describes the usefulness of Norine to discover novel nonribosomal peptides, with examples of biological results obtained thanks to Norine tools. More than 20 NRPs have already been submitted since September, proving the efficiency of our communication and usefulness and relevancy of Norine.
- **Monomeric structure.** The tool Smiles2Monomers (abbreviated s2m) infers efficiently and accurately the monomeric structure of a polymer from its chemical structure [1]. It is provided to the scientific community through the Norine website for on-line run or for download. Beside its utility to facilitate the annotation of new peptides, it allowed us to detect annotation errors in the Norine database.

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

High-throughput V(D)J repertoire analysis is an activity started in the group in 2012. As mentioned in previous reports, we produced a platform dedicated to analysing lymphocyte populations: Vidjil. Starting from DNA sequences, Vidjil is able to identify and quantify lymphocyte populations, visualise them and store metadata. Vidjil is now used routinely in Lille hospital and is also tested in other laboratories around the world.

With collaborators in Prague we used Vidjil in a retrospective study on patients suffering acute lymphoblastic leukemia [3]. The study identified a new measure of predicting relapse in patients, just a month after the diagnosis. This measure is simple as it relies on the diversity of the lymphocyte population.

7.4. Spaced seed coverage

In the field of spaced seed statistics these last two years, a new challenge is the selection of a set of spaced seeds that are at the same time sensitive, while providing a stable similarity measure for *alignment-free genomic sequence comparison*. One of the most stable estimators is the *coverage* provided by these seeds. We have proposed an efficient method to build the coverage automaton, in order to compute several statistics efficiently. This work was implemented in Iedera and published in the AISM journal [4].

7.5. Genome scaffolding with contaminated data

Scaffolding is a cornerstone in the assembly of genomes from next-generation sequencing data. It consists in ordering assembled sequences according to their putative order and orientation in the source genome. However, we are almost always in a setting where the genome is not known. Instead, order and orientation of sequences are inferred from partial information present in the sequencing data.

Unfortunately, sequencing data is noisy and often has contamination, i.e. a subset of the data which indicates a wrong genome order and/or orientation. We have investigated this effect and designed the first algorithm that explicitly models this contamination to better perform scaffolding.

This work appeared in the proceedings of the WABI 2015 conference [9] and has been accepted to the Bioinformatics journal, currently under revision. This work is in collaboration with K. Sahlin and L. Arvestad (KTH, Sweden).

7.6. Mining metatranscriptomic data

The team has recently developed the SortMeRNA software, which is a sequence analysis tool for filtering, mapping and OTU-picking NGS reads. The core algorithm is based on approximate seeds and allows for fast and sensitive analyses of nucleotide sequences. In [11], we demonstrate a computational technique for filtering ribosomal RNA from total RNA in metatranscriptomic data using it. Additionally, we propose a post-processing pipeline using the latest software tools to conduct further studies on the filtered data, including the reconstruction of mRNA transcripts for functional analyses and phylogenetic classification of a community using the ribosomal RNA. This work is a collaboration with Genoscope.

7.7. Structured RNAs

In many families of strutured RNAs, the signature of the family cannot be characterized by a single consensus structure, and is mainly described by a set of alternate secondary structures. For example, certain classes of RNAs adopt at least two distinct stable folding states to carry out their function. This is the case of riboswitches, that undergo structural changes upon binding with other molecules, and recently some other RNA regulators were proven to show evolutionary evidence for alternative structure. The necessity to take into account multiple structures also arises when modeling an RNA family with some structural variation across species, or when it comes to work with a set of predicted suboptimal foldings. In this perspective, we have introduced the concept of RNA multistructures, that is a formal grammar based framework specifically designed to model a set of alternate RNA secondary structures. Continuing our work of 2014, we provide several motivating examples and propose an efficient algorithm to search for RNA multistructures within a genomic sequence. This work was published in [7].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

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

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR

- 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 increase 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 the installation, administration and usability. This will make the software well suited for a daily clinical use. The software is already used in test on our own web server (more than 1,000 samples processed by now). Our goal is that several labs use Vidjil on a daily basis by the end of the ADT, and that they all have their own Vidjil server.

9.1.3. Others

- PEPS Gen-CoV: *Global bioinformatics analysis of coronavirus strain 229E in hospital outbreak.* The goal of this PEPS is to provide with a better characterization of coronavirus infections and to understand underlying mecanisms that lead to the high diversity of coronaviruses. To achieve this goal, we will sequence and analyze a number of coronavirus 229E genomes in order to characterize their diversity, identify features that influence pathogenicity and propose a model of evolution. All those results will be correlated with epidemiologic data thanks to a partnership with Lille hospital.
- PEPS JCJC: *Frugal algorithms for third-generation DNA sequencing*. The goal of this PEPS is to develop lightweight algorithms and data structures for the analysis of third-generation sequencing data. Among third-generation technologies, the MinION sequencer is a new, portable USB device that can perform DNA sequencing using only common lab equipment and a laptop computer. However, analysis of the data produced by the MinION can only be carried by uploading data to a cloud server. Indeed, all algorithms and data structures that are currently known require large computational resources to process such data. This is unfortunate for at least two reasons: analysis of the data now takes more time than its production, and confidential data needs to be processed on potentially insecure cloud servers. We seek to design methods that would enable analysis of sequenced data on the same machine as the one that performed sequencing.

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*). The subject is 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.

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, informatique et mathématique (LaCIM) - Anne Bergeron

Start year: 2014

See also: http://thales.math.uqam.ca/~cgalcode/

The aim of this Associated Team is the development of comparative genomics models and methods for the analysis of eukaryotes gene structure evolution. The goal of the project 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 meetings of CG-ALCODE researchers took place in 2015. First, Samuel Blanquart, Anne Bergeron and Krister Swenson met each other in Montpellier, from 27th to 30th of April. Second, Samuel Blanquart, Jean Stéphane Varré spent two weeks in Montréal, from 1st to 11th November, to work with Anne Bergeron.

9.3.1.2. Informal International Partners

- *Astrid Lindgrens Hospital, Stockholm University:* Collaboration with Anna Nilsson and Shanie Saghafian-Hedengren on RNA sequencing of stromal cells.
- *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.
- *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.
- *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 and Markus Mueller on nonribosomal peptides.
- *School of Social and Community Medicine, University of Bristol:* Collaboration with John Moppett on leukemia follow-up.
- *Science for Life Laboratory, Stockholm University:* Collaboration with Lars Arvestad and Kristoffer Sahlin on genome scaffolding of contaminated libraries.
- *Theoretical Biochemistry Group, Universität Wien:* Collaboration with Andrea Tanzer and Ronny Lorenz on RNA folding and RNA kinetics.

9.4. International Research Visitors

9.4.1. Visits of International Scientists

- Kristina Heyn, PhD student, Institut für Biophysik und physikalische Biochemie, University of Regensburg (from 6th to 11th of July)
- Burkhard Morgenstern, professeur, Universität Gottingen (from 20th to 23th of April)
- Samarth Rangavittal, PhD student, The Pennsylvania State University (from October 18th to December 6th)
- Gabriele Valiente, professeur, Universitat Politècnica de Catalunya (from 25th to 29th of May)
- Tilmann Weber, senior researcher, Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark (from 18th of October to 31st of October)

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. Member of the organizing committees

- workshop *3rd generation sequencing*, June 9, 117 attendees
- workshop *Bioinformatics tools for NRPS discovery, from genomic data to the products*, october 28-30, 32 attendees
- workshop *Drug design and virtual screening* November 30, 60 attendees

10.1.2. Scientific events selection

10.1.2.1. Chair of conference program committees

- H. Touzet was PC chair of the international conference WABI 2015 (Atlanta, USA, september 10-12)

10.1.2.2. Member of the conference program committees

- HiCOMB 2015 (M. Giraud)
- JOBIM 2015 (H. Touzet).
- RECOMB-seq 2015 (H. Touzet, L. Noé).
- SeqBio 2015 (H. Touzet).
- WABI 2015 (H. Touzet, L. Noé, R. Chikhi).

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing activities

- Bioinformatics (M. Pupin, L. Noé, R. Chikhi)
- BMC Bioinformatics (M. Salson, R. Chikhi)
- Genome Biology (R. Chikhi)
- GigaScience (R. Chikhi)
- Microbiology Open (V. Leclère)
- Nucleic Acids Journal (R. Chikhi)
- Synthetic and Systems Biotechnology (M. Pupin)

10.1.4. Scientific expertise

- Expert for Allistène (H. Touzet)
- Expert for the French-Brazilian program CAPES-COFECUB (H. Touzet)
- Expert for Département de l'Essone (H. Touzet)

10.1.5. 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 national evaluation committee of computer science for University members (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 Gilles Kahn PhD award national commitee (H. Touzet)
- Member of the CRIStAL Laboratory council (H. Touzet)
- Member of the GRIOTE (Groupement de Recherche en Intégration de données Omics à Très grande Echelle, région Pays de Loire) scientific council (H. Touzet)
- Member of the scientific committee of the national program Environmics (H. Touzet)
- Member of the CRIStAL scientific council, representant of the thematic group “Modeling for life sciences” (J.-S. Varré)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching administration

- 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)
- 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 licence semester "Computer Science – S3 Harmonisation (S3H)", univ. Lille 1 (L. Noé)

10.2.2. Teaching

Licence: 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: Y. Dufresne, *Algorithmics and complexity*, 36h, M1 Computer Science, Univ. Lille 1

Master: M. Giraud, *Algorithms for RNA Analysis*, 12h, M2 Bioinformatique et Modélisation, Univ. Paris 6

Licence: 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

Master: V. Leclère, *Bioinformatics*, 30h, M2 Transformation Valorisation Industrielles des Agro-ressources, Univ. Lille 1

Master: V. Leclère, *Secondary metabolites*, 20h, M1 Biology, Univ. Lille 1

Master: V. Leclère, *Biotechnology*, 20h, M1Biology, Univ. Lille 1

Master: V. Leclère, *Microbiology*, 20h, M1 Biology, Univ. Lille 1

Master: L. Noé, *Bioinformatics*, 40h, M1 Biotechnologies, 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: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:M. Pupin *Introduction to programming (Python)*, 36h, L1 Computer science, Univ. Lille 1

License:M. Pupin, *Databases*, 36h, L3 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

Master: M. Pupin, M. Salson *Bioinformatics*, 40h, M1 Biology and Biotechnologies, Univ. Lille 1

License:T. Rocher, *Algorithmics and programming*, 28h, L3 Polytech'Lille, Univ. Lille 1

License:T. Rocher, *Algorithmics and programming (remedial course)*, 14h, L3 Polytech'Lille, Univ. Lille 1

License:T. Rocher, *Databases*, 24h, L3 Polytech'Lille, Univ. Lille 1

License: M. Salson, *Automata and language theories*, 36h, L3 Computer science, Univ. Lille 1

License: M. Salson, *Skeptical thinking*, 30h, L3 Computer science, Univ. Lille 1

License: M. Salson, *Coding and information theory*, 63h, L2 Computer science, Univ. Lille 1

Master: M. Salson, *Skeptical thinking*, 14h, M2 Journalist and Scientist, ESJ, Univ. Lille 1

Master: M. Salson, *Algorithms for life sciences*, 18h, M2 Complex models, algorithms and data, Univ. Lille 1

License: J.-S. Varré, *Web programming*, 36h, L2 (licence “Computer Science”, Univ. Lille 1

License: J.-S. Varré, *Programming with Python*, 36h, L2 (licence “Sciences for Engineers”, Univ. Lille 1

License: J.-S. Varré, *Algorithms and Data structures*, 50h, L2 (licence “Computer science”, Univ. Lille 1

License: J.-S. Varré, *System*, 36h, L3 Computer science”, Univ. Lille 1

10.2.3. Supervision

PhD in progress: Y. Dufresne, Modèles et algorithmes pour la gestion de la biodiversité des peptides non-ribosomiques et la mise en évidence de nouveaux peptides bioactifs, 2013/10/01, 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, J.-S. Varré, 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: C. Vroland, Indexing data for microRNA and microRNA target site identification in genomes, 2012/10/01, H. Touzet, V. Castric (EEP), M. Salson

10.2.4. Juries

- Member of the thesis committee of Souhir Sabri (CIRAD, Montpellier, V. Leclère)
- Member of the PhD thesis jury of Laetitia Bourgeade (LABRI, Bordeaux, H. Touzet)
- Member of the thesis committee of Florian Plaza Onate (Ecole Centrale, Paris, R. Chikhi)
- Member of the thesis committee of Jérome Audoux (INSERM, Montpellier, R. Chikhi, M. Salson)
- Member of the HDR jury of Fabien Jourdan (INRA, Toulouse, H. Touzet)
- Member of the hiring committee Professor of Université de Nantes (H. Touzet)
- Member of the hiring committee Professor of Université d’Evry (H. Touzet)
- Member of the hiring committee McF of Université Lille 1 (H. Touzet)
- Member of the hiring committee Professor of University of Montréal (H. Touzet)
- Member of the hiring committee ATER of Université Lille 1 (L. Noé, J.S. Varré)

10.3. Popularization

- We made seven presentations, using dedicated “genome puzzles” in high schools during the “Science week” to popularize bioinformatics.
- During a whole day in June we made presentations on bioinformatics with our “genome puzzles” to several groups of high school students.
- V. Leclère has created a demonstration stand for Xperium, part of the Learning center Innovation of Lille 1 University.

11. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] 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

[2] 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

[3] M. Kotrova, K. Muzikova, E. Mejstrikova, M. Novakova, V. Bakardjieva-Mihaylova, K. Fiser, J. Stuchly, M. Giraud, M. Salson, C. Pott, M. Brüggemann, M. Füllgrabe, J. Stary, J. Trka, E. Fronkova. *The predictive strength of next-generation sequencing MRD detection for relapse compared with current methods in childhood ALL*, in "Blood", August 2015, vol. 126, n^{o} 8, pp. 1045-7 [*DOI :* 10.1182/blood-2015-07-655159], https://hal.archives-ouvertes.fr/hal-01241663

[4] 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

[5] A. Perrin, J.-S. Varré, S. Blanquart, A. Ouangraoua. *ProCARs: Progressive Reconstruction of Ancestral Gene Orders*, in "BMC Genomics", 2015, vol. 16, n^{o} Suppl 5, S6 p. [*DOI :* 10.1186/1471-2164-16-S5-S6], https://hal.inria.fr/hal-01217311

[6] 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

[7] A. Saffarian, M. Giraud, H. Touzet. *Modeling alternate RNA structures in genomic sequences*, in "Journal of computational biology : a journal of computational molecular cell biology", February 2015, vol. 22, n^{o} 3, pp. 190-204, https://hal.archives-ouvertes.fr/hal-01228130

Invited Conferences

[8] T. T. Tran, M. Giraud, J.-S. Varré. *Perfect Hashing Structures for Parallel Similarity Searches*, in "International Workshop on High Performance Computational Biology (HiCOMB 2015) / International Parallel and Distributed Processing Symposium (IPDPS 2015)", Hyderabad, India, 2015, pp. 332-341 [*DOI :* 10.1109/IPDPSW.2015.105], https://hal.inria.fr/hal-01153893

International Conferences with Proceedings

[9] K. Sahlin, R. Chikhi, L. Arvestad. *Genome scaffolding with PE-contaminated mate-pair libraries*, in "WABI 2015", Atlanta, United States, 2015, http://biorxiv.org/content/early/2015/08/28/025650 [*DOI :* 10.1101/025650], https://hal.archives-ouvertes.fr/hal-01236176

Conferences without Proceedings

[10] I. GUIGON, S. LEGRAND, J.-F. BERTHELOT, M. BENMOUNAH, H. TOUZET. *Finding and analysing microRNAs in plant genomes with miRkwood*, in "JOBIM 2015", Clermont-Ferrand, France, July 2015, https://hal.archives-ouvertes.fr/hal-01247604

Scientific Books (or Scientific Book chapters)

[11] E. KOPYLOVA, L. NOÉ, C. DA SILVA, J.-F. BERTHELOT, A. A. ALBERTI, J.-M. AURY, H. TOUZET. *Deciphering metatranscriptomic data*, in "Methods in Molecular Biology", RNA Bioinformatics, Springer, January 2015, vol. 1269, pp. 279-291 [*DOI :* 10.1007/978-1-4939-2291-8_17], https://hal.inria.fr/hal-01104015

Books or Proceedings Editing

[12] M. POP, H. TOUZET (editors). *Algorithms for bioinformatics*, 15th International Workshop, WABI 2015, Atlanta, GA, USA, September 10-12, 2015, Proceedings, Springer Verlag, Atlanta, United States, September 2015, n^{o} 9289 [*DOI :* 10.1007/978-3-662-48221-6], https://hal.archives-ouvertes.fr/hal-01228138