IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2014

Project-Team BONSAI

Bioinformatics and Sequence Analysis

IN COLLABORATION WITH: Laboratoire d'informatique fondamentale de Lille (LIFL)

RESEARCH CENTER
Lille - Nord Europe

THEME
Computational Biology

Table of contents

Project-Team BONSAI

Keywords: Computational Biology, Genomics, Next Generation Sequencing, Rna Annotation, Nonribosomal Peptides, Genome Rearrangement

Creation of the Project-Team: 2011 January 01.

1. Members

Research Scientists
- Hélène Touzet [Team leader, CNRS, Senior Researcher, HdR]
- Samuel Blanquart [Inria, Researcher]
- Rayan Chikhi [CNRS, from Nov 2014]
- Mathieu Giraud [CNRS, Researcher]
- Aïda Ouangraoua [Inria, Researcher, until Sep 2014]

Faculty Members
- Stéphane Janot [Univ. Lille 1, Associate Professor]
- Valerie Leclère [Univ. Lille 1, Associate Professor, from Sep 2014, HdR]
- Laurent Noé [Univ. Lille 1, Associate Professor]
- Maude Pupin [Univ. Lille 1, Associate Professor, HdR]
- Mikaël Salson [Univ. Lille 1, Associate Professor]
- Jean-Stéphane Varré [Univ. Lille 1, Professor, HdR]

Engineers
- Thierry Barthel [Inria, until Oct 2014]
- Jean-Frédéric Berthelot [CNRS, until Jun 2014]
- Marc Duez [Univ. Lille 2, SIRIC OncoLille]
- Areski Flissi [CNRS]
- Isabelle Guigon [CNRS, from Sep 2014]
- Alan Lahure [CNRS]
- Juraj Michalik [CNRS, from Oct 2014]
- Amandine Perrin [Inria, from Mar 2014]

PhD Students
- Yoann Dufresne [PhD, Univ. Lille 1]
- Pierre Pericard [PhD, Univ. Lille 1]
- Tatiana Rocher [PhD, Univ. Lille 1, from Nov 2014]
- Chadi Saad [PhD, Univ. Lille 2, from Oct 2014]
- Léa Siegwald [PhD, Univ. Lille 2, from Mar 2014]
- Christophe Vroland [PhD, Univ. Lille 1]

Visiting Scientists
- Anne Bergeron [UQÀM, Jul 2014]
- Paul Guertin [UQÀM, Jul 2014]

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. Most of our research 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. Combinatorial discrete models and algorithms

Our research is driven by biological questions. At the same time, 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. High-performance computing is another tool that we use to achieve our goals.

3.2. Discrete statistics and probability

At a lower level, our work relies on a basic background on discrete statistics and probability. When dealing with large input data sets, it is essential to be able to discriminate between noisy features observed by chance from those that are biologically relevant. The aim here is to introduce a probabilistic model and to use sound statistical methods to assess the significance of some observations about these data. Examples of such observations are the length of a repeated region, the number of occurrences of a motif (DNA or RNA), the free energy of a conserved RNA secondary structure, etc. Probabilistic models are also used to describe genome evolution. In this context, Bayesian models and MCMC sampling allow to approximate probability distributions over free parameters and to describe biologically relevant models.

4. Application Domains

4.1. 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, ...

4.2. 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.

4.3. 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.

4.4. 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 [14]. 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).

5. New Software and Platforms

5.1. SortMeRNA – Metatranscriptome classification

Software web site: http://bioinfo.lille.inria.fr/RNA/sortmerna

Licence: GPL

Objective: *SortMeRNA* is a tool designed to rapidly filter ribosomal RNA fragments from metatranscriptomic data produced by next-generation sequencers. It is available for download from our website, or through the open web-based platform Galaxy. The development version is also available on GitHub. *SortMeRNA* was first released in October 2012. It is now used in production by Genoscope (French National Center for Sequencing) to process all metatranscriptomic data of the Tara Ocean Expedition, and has been integrated in several other computational pipelines (Qiime developed at University of Colorado at Boulder, MetaMetadb developed at University of Tokyo, Leimena pipeline developed at Wageningen University,...).

SortMeRNA is still under development through a partnership with the Knight lab (University of Colorado at Boulder). Version 2.0 has been released in November 2014, and has extended functionalities. It can now perform sequence alignments to any ribosomal RNA database, which allows the user to study the taxonomic content of a microbial sample. This new version has been presented at the international workshops [12], [11].

5.2. Vidjil – Quantifying lymphocyte rearrangements in high-throughput sequencing data

Software web site: http://bioinfo.lille.inria.fr/vidjil/

Objective: **Vidjil** is a platform for high-throughput V(D)J recombinations analysis. containing three components. The Vidjil *algorithm* process high-througput sequencing data to extract V(D)J junctions and gather them into clones. Vidjil starts from a set of reads and detects “windows” overlapping the actual CDR3. This is based on an fast and reliable seed-based heuristic and allows to output all sequenced clones. The analysis is extremely fast because, in the first phase, no alignment is performed with database germline sequences [5]. The Vidjil dynamic *browser* is made for the visualization and analysis of clones and their tracking along the time in a “minimal residual disease” setup or in a immunological study. The browser visualize data processed by the Vidjil algorithm or by other V(D)J analysis pipeline and enables to explore further cluterings. Finally, a *patient database* with a server links the browser and the algorithmic part. The goal is that the clinicians will be able to upload, manage and process their runs on a server hosted in their hospital.

In 2014, the development of Vidjil was supported by the SIRIC OncoLille (Marc Duez). We developed the new patient database and added features both on the browser and on the algorithm (multi-system analysis). Several hospital labs in France and in Europe are testing Vidjil. The Lille hospital plans to use Vidjil in 2015 in a pre-production pipeline.

5.3. Norine – A resource for nonribosomal peptides

Software web site: http://bioinfo.lille.inria.fr/norine/

Objective: **Norine** is a public computational resource that contains a database of NRPs with a web interface and dedicated tools, such as a 2D graph viewer and editor for peptides or comparison of NRPs. Norine was created and is maintained by members of BONSAI team, in tight collaboration with members of the ProBioGEM lab, a microbial laboratory of Lille1 University. Since its creation in 2006, Norine has gained an international recognition as the unique database dedicated to non-ribosomal peptides because of its high quality and manually curated annotations, and has been selected by wwPDB as a reference database. It is queried from all around the world by biologists or biochemists. It receives more than 3000 queries per month.

To enhance the Norine resource, we have recently developed a new module, named MyNorine, which is an open interface for biologists and biochemists dedicated to the submission of new non-ribosomal peptides in Norine database. Up to now, peptides were manually inputted and verified before being added in the database, which could potentially lead to human errors. The goal of MyNorine is to help users during the submission of peptides and monomers, by guiding them during all steps. For that, users, all over the world, can create an account on MyNorine. Thus, they contribute to the Norine resource and become curators (author of a peptide entry is mentioned in the corresponding page of Norine). Submitted peptides/monomers are validated, through a workflow process, by Norine team members, to ensure correct and consistent entries.

5.4. miRkwood –microRNAs in plant genomes

Software web site: http://bioinfo.lille.inria.fr/mirkwood/

Objective: **miRkwood** is a web server for the identification of hairpin precursors of both conserved and non-conserved miRNAs in plant genomes. It is able to face the diversity of plant pre-miRNAs and is optimised to take advantage of their distinctive properties: Sequence length, secondary structure, free energy, miRNA conservation, stability of the miRNA/miRNA* duplex, ... Moreover, it offers an intuitive and comprehensive user interface to navigate in the data, as well as many export options to allow the user to conduct further analyses on a local computer. Ongoing work is concerned with integrating small RNA-seq data.

5.5. ProCARs

Software web site: http://bioinfo.lille.inria.fr/procars

Objective: **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. The method output CARs as sets of ordered contiguous blocks in the targeted ancestor. ProCARs has been developped with Python 2.7.5.

6. New Results

6.1. Highlights of the Year

- Amandine Perrin received the best paper award and the best oral presentation at the ISCB-LA 2014 international conference for the work on reconstruction of ancestral gene orders.
- Hélène Touzet was invited as a keynote speaker at the ALGO 2014 international conference. The topic of the talk was RNA bioinformatics.

BEST PAPER AWARD :
[7] **ProCARs: Progressive Reconstruction of Ancestral Gene Orders in ISCB-Latin America**. A. PERRIN, J.-S. VARRÉ, S. BLANQUART, A. OUANGRAOUA.

6.2. High-throughtput sequence processing

- **Analysis of immunological rearrangements for leukemia diagnosis and monitoring.** High-throughput sequencing is spreading in the hospitals and many classical routines are now being transferred to this new technology. However in the specific case of lymphocyte monitoring, some complications arise. Classical bioinformatics software tools do not apply to the specificity of lymphocyte rearrangements. That is why we developed the software Vidjil (see 5.2) together with Lille hospital. This work has been published [5] and was also presented, as a poster, during the annual conference of the American Society of Hematology (ASH) [13]. We are now members of the EuroClonality-NGS work group which aims at providing a standardized way of monitoring leukemia using high-throughput sequencing at the European level.

- **New seeds for approximate pattern matching.** We addressed the problem of approximate pattern matching using the Levenshtein distance. Given a text T and a pattern P, find all locations in T that differ by at most k errors from P. For that purpose, we proposed a filtration algorithm that is based on a novel type of seeds, combining exact parts and parts with a fixed number of errors, that we called 01^*0 seed. Implementation has been performed on a Burrows-Wheeler transform. Experimental tests show that the method is specifically well-suited to search for short patterns (< 50 letters) on a small alphabet (*e.g.* DNA alphabet) with a medium to high error-rate (7 %–15 %). This work has been published in [9], and has a large number of applications in computational biology, such as finding microRNA targets, for example.
- **Spaced seeds and Transition seeds.** This year, two collaborative works have been published on the topic of spaced seeds and derivated models. The first work, resulting from a collaboration with Martin C. Frith from the *Computational Biology Research Center* (Tokyo), increases the sensivity of several search tools (among them, LAST, LASTZ, YASS,...) by computing specific seeds adapted to transition ratios observed during Eucaryotic comparisons. This work has been published in [3], together with the design of seeds obtained. The second work, issued from collaboration with Donald E.K. Martin from the *Department of Statistics* of the *North Carolina State University* (Raleigh), deals with the coverage of spaced seeds and shows how this criterion helps selecting good seeds for SVM string-kernels and alignment-free distances. This work has been published in [6].

6.3. RNA algorithms

- **A universal framework for RNA algorithms.** We have proposed a new generic specification framework, called *inverted coupled rewrite systems* that can deal with optimization problems on strings, trees, and arc-annotated sequences. It is specifically well-suited to handle RNA algorithms, such as alignment or folding algorithms. It is based on the following ideas. The solutions of combinatorial optimization problems are the inverse image of a term rewrite relation that reduces problem solutions to problem inputs. A tree grammar is used to further refine the search space, and optimization objectives are specified as interpretations of these terms. All these constituents provide a mathematically precise and complete problem specification, leading to concise yet translucent specifications of dynamic programming algorithms. This work is a collaborative project with R. Giegerich from Universität Bielefeld, and has been published in [4].
- **RNA multistructures.** In many RNA families, 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. We provide several motivating examples and propose an efficient algorithm to search for RNA multistructures within a genomic sequence. This work was published in [8].

6.4. 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. We designed a progressive homology-based method which iteratively detects and assembles ancestral adjacencies while allowing some micro-rearrangements of synteny blocks at the extremities of the progressively assembled CARs. Comparing with other methods we are able to produce more robust CARs with a very simple and efficient method. This work was published in [7].

6.5. Nonribosomal peptides

- **Monomeric structure.** The algorithm that identifies the monomeric structure of a polymer from its chemical structure has been finished and named s2m. It is based on a double index: A partial index constructed on the monomer database that uses a markovian model to speed up the search time ; and an index constructed on the fly on the studied polymer. This strategy was originally developed for nonribosomal peptides, but can be applied to any polymer.
- **Florine: Nonribosomal peptide synthetase annotations.** Florine [2] is a workflow dedicated to the discovery of new nonribosomal peptide synthetases. It describes sequential steps starting from DNA sequences leading to the design of candidate bioactive peptides. It is a useful tool for new drug discovery because it can be applied whatever the producing micro-organisms as it takes into account the enzymatic specificities related to each genus. This work was performed in collaboration with members of EPI Orpailleur (CRI Nancy Grand Est), Marie-Dominique Devignes and Malika Smaïl-Tabbone.
- **Activity prediction of small molecules.** Bayesian Belief Network was used for the first time to classify compounds according to their biological activity [1]. This method was applied on nonribosomal peptides and gave promising results on predicting their activity.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

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

8. Partnerships and Cooperations

8.1. Regional Initiatives

- Projet émergent call 2011. “Scénarios d’évolution génomique basés sur les régions de cassure des réarrangements génomiques” involving GEPV (UMR CNRS 8198, Université Lille 1) and BONSAI. The project led to the recruitment of Amandine Perrin in 2014.
- SIRIC OncoLille supports our research in collaboration with Lille hospital on quantification of lymphocyte rearrangements, funding the contract of Marc Duez in 2014.

8.2. National Initiatives

8.2.1. 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 platforms 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
- Mastodons (2014): National funding from CNRS (call *Scientific big data*). This call targets the management, analysis and exploitation of massive scientific data sets. We have a collaborative project for Next Generation Sequencing data analysis with LIRMM (Montpellier) and Genscale (Inria Rennes).

- PEPS Bio-Math-Info *ReSeqVar* (2013-2014): National funding from CNRS. This new project aims at designing new read mapping algorithms in the context of human genome resequencing, taking into account known variants. There are two partners: UMR 8199 (Génomique et maladie métabolique, Ph Froguel, O. Sand, part of the LIGAN sequencing platform) and BONSAI.

8.2.2. ADT

- ADT biosciences resources (2012-2014): This ADT aims to build a portal of available applications in bioinformatics at Inria. The projects involves all the 8 teams from theme Bio-A and is more specifically developed by BONSAI and Rennes. The engineer hired from 2012 to 2014 in Lille finished its contract at fall. The portal is available at http://ibr.genouest.org.

8.3. European Initiatives

8.3.1. 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.
- EuroClonality-NGS: This working group belongs to the ESLHO (European Scientific foundation for Laboratory HematoOncology), which aims at standardizing laboratory diagnostics focused on lymphoid malignancies, it is also responsible for quality controls of European laboratories. The EuroClonality-NGS working group itself is dedicated to provide new standards using high-throughput sequencing.

8.4. International Initiatives

8.4.1. Inria Associate Teams

8.4.1.1. CG-ALCODE

The title of the project is “Comparative Genomics for the analysis of gene structure evolution: ALternative CODing in Eukaryote genes through alternative splicing, transcription, and translation.”. The project involves partners from EPI BONSAI and from the Université du Québec À Montréal (UQÀM, Canada), from year 2014 to year 2017 (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 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.

8.4.2. Inria International Partners

8.4.2.1. Informal International Partners

- *Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark:* Collaboration with Tilmann Weber on nonribosomal peptides.
- *Computational Biology Research Center, Tokyo*: Collaboration with Martin C. Frith on transition spaced seeds [3].
- *Department of Statistics of the North Carolina State University (Raleigh)*: Collaboration with Donald E.K. Martin one spaced seeds coverage [6].
- *Institut für Biophysik und physikalische Biochemie’, University of Regensburg*: Collaboration with Rainer Merkl on ancestral sequence inference and synthesis.
- *University of Bielefeld*: Collaboration with Robert Giegerich on RNA bioinformatics [4].

8.5. International Research Visitors

8.5.1. Visits of International Scientists

- Anne Bergeron, professor, UQÀM, Canada (from July 7 to July 11 2014).
- Paul Guertin, UQÁM (from July 7 to July 24).

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. general chair, scientific chair

- Proteomics for omics analysis, 2014/3/2, 35 people
- Integrative biology, 2014/17/12, 60 people

9.1.1.2. member of the organizing committee

- Polaris seminar series (M. Giraud)

9.1.2. Scientific events selection

9.1.2.1. member of the conference program committee

- RECOMB-seq (L. Noé, H. Touzet).
- ECCB'14 poster session (H. Touzet).
- SeqBio 2014 (H. Touzet).

9.1.2.2. reviewer

- Laurent Noé: AFL 2014.

9.1.3. Journal

9.1.3.1. reviewer

- Algorithms for Molecular Biology (L. Noé), Acta Biotheoretica (L. Noé), Bioinformatics (H. Touzet)

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

L. Noé, *Bioinformatics*, 54h, M1 (master "Génomique Protéomique", univ. Lille 1)

L. Noé, *Networks*, 42h, L3 (licence "Computer science", univ. Lille 1)

M. Pupin, M. Salson, *Introduction to programming (OCaml)*, 96h, L1 (licence "Computer science", univ. Lille 1)

M. Salson, *Coding and information theory*, 63h, L2 (licence "Computer science", univ. Lille 1)

J.-S. Varré *Web programming*, 36h, L2 (licence "Computer Science", univ. Lille 1)

J.-S. Varré *Programming with Caml*, 36h, L2 (licence "Sciences for Engineers", univ. Lille 1)

J.-S. Varré *Algorithms and Data structures*, 50h, L2 (licence "Computer science", univ. Lille 1)

J.-S. Varré, *System*, 36h, L3 (licence "Computer science", univ. Lille 1)

M. Salson, *C programming*, 42h, L3 (licence "Computer science", univ. Lille 1)

M. Pupin, M. Salson *Bioinformatics*, 100h, M1 (master "Biology and Biotechnologies", univ. Lille 1)

S. Blanquart, *Algorithms and applications in bioinformatics*, 24h, M1 (master "Computer Science", univ. Lille 1)

V. Leclère *Bioinformatics*, 50h, M2 (master "Transformation Valorisation Industrielles des Agro-ressources", univ. Lille 1)

M. Pupin, J.-S. Varré *Computational biology*, 30h, M2 (master "Modèles complexes, algorithmes et données", univ. Lille 1)

S. Blanquart, *Methods in phylogenetics*, 4h, M2 (master "Ecology Environment", univ. Lille 1)

M. Giraud, *Algorithms for RNA Analysis*, 12h, M2 (master "Bioinformatique et Modélisation", univ. Paris 6)

M. Salson, *Skeptical thinking*, 14h, M2 (master "journalist and scientist", univ. Lille1, ESJ)

J.-S. Varré, *ISN - Computer science for secondary school*, 20h, second-level teachers.

9.2.2. Supervision

PhD in progress: C. Vroland, Indexing data for microRNA and microRNA target site identification in genomes, 2012/10/01, H. Touzet, V. Castric (GEPV), M. Salson

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: P. Pericard, Methods for taxonomic assignation in metagenomics, 2013/11/01, H. Touzet, S. Blanquart.

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: 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: L. Siegwald, Bionformatic analysis of Ion Torrent metagenomic data, 2014/01/03, H. Touzet, Y. Lemoine (Institut Pasteur de Lille)

9.2.3. Juries

- Member of the thesis committee of Jérémie Gilliot (Université Lille 1, J.-S. Varré)
- Member of the thesis committee of Trong-Tuan Vu (Université Lille 1, J.-S. Varré)
- Member of the thesis committee of Audrey Vingadassalon (Université Paris Sud, V. Leclère)
- Member of the thesis committee of Safwan Saker (Université de Toulouse, V. Leclère)
- Member of the thesis committee of Valentin Wücher (Université Rennes 1, H. Touzet)
- Member of the thesis committee Susan Higashi (Université Lyon 1, H. Touzet)
- Member of the thesis committee of Mateusz Pawlik (Freie Universität Bozen, H. Touzet)
- Member of the tenure committee of Sylvie Hamel (UQAM, H. Touzet)
- Member of the national Inria CR1 hiring committee (M. Giraud)
- Member of the CR2 hiring committee of Inria Rocquencourt (M. Giraud)
- Member of the McF 27MC1727 committee (Université Pierre et Marie Curie, L. Noé)
- Member of the McF 27/64, poste 0916 committee (Université de Reims, M. Pupin)
- Member of the hiring committee McF of Université Lyon 1 (H. Touzet)

9.2.4. Administrative activities

- National representative (*chargée de mission*) for the Institute for Computer Sciences (INS2I) in CNRS [1]. She is more specifically in charge of relationships between the Institute and life sciences (H. Touzet)
- Member of the Inria evalution commitee (M. Giraud)
- Member of the Inria local committee for scientific grants (A. Ouangraoua)
- Member of the Inria local committee for technology development (M. Pupin)
- Member of the Gilles Kahn PhD award national commitee (H. Touzet)
- Member of the national ANR evaluation committe CES19 (H. Touzet)
- Member of CSS MBIA (mathematics, bioinformatics and artificial intelligence) at INRA (H. Touzet)
- Member of the executive council of the IFB, Institut Français de Bioinformatique, (M. Pupin)
- Member of CUB Inria Lille Nord Europe (S. Blanquart).
- Head of PPF bioinformatics – University Lille 1 (H. Touzet)
- Head of Bilille, Lille bioinformatics platform (M. Pupin)
- Head of IFB-NE (pôle Nord-Est de l'Institut Français de Bioinformatique), a cluster of 4 bioinformatics platforms (M. Pupin)
- Member of UFR IEEA council (M. Pupin)
- Member of UFR Biologie council (V. Leclère)
- Head of the GIS department (Statistics and Computer Sciences) of Polytech'Lille (S. Janot)
- Head of the master "Transformation Valorisation Industrielles des Agro-ressources", univ. Lille 1 (V. Leclère)
- Head of the master "Modèles complexes, algorithmes et données", univ. Lille 1 (J.-S. Varré)
- Member of the ProBioGEM Laboratory council (V. Leclère)
- Member of the ProBioGEM scientific committee (V. Leclère)
- Member of the scientific operational committee of Xperium, Univ. Lille 1 (V. Leclère)
- Member of the LIFL Laboratory council (L. Noé, H. Touzet)
- Head of the thematic group "Modeling for Life Sciences" of CRIStAL and member of the scientific council of CRIStAL (J.-S. Varré)

9.3. 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.

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] 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° 1, pp. 30-36 [*DOI :* 10.1021/CI4004909], https://hal.archives-ouvertes.fr/hal-01090611

[1] CNRS: National Center for Scientific Research

[2] T. CARADEC, M. PUPIN, A. VANVLASSENBROECK, M.-D. DEVIGNES, M. SMAÏL-TABBONE, P. JACQUES, V. LECLÈRE. *Prediction of Monomer Isomery in Florine: A Workflow Dedicated to Nonribosomal Peptide Discovery*, in "PLoS ONE", January 2014, vol. 9, n° 1, e85667 [*DOI :* 10.1371/JOURNAL.PONE.0085667], https://hal.archives-ouvertes.fr/hal-01090619

[3] 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 [*DOI :* 10.1093/NAR/GKU104], https://hal.inria.fr/hal-00958207

[4] 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

[5] 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

[6] 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", December 2014, vol. 21, n° 12, 28 p. [*DOI :* 10.1089/CMB.2014.0173], https://hal.inria.fr/hal-01083204

International Conferences with Proceedings

[7] *Best Paper*
A. PERRIN, J.-S. VARRÉ, S. BLANQUART, A. OUANGRAOUA. *ProCARs: Progressive Reconstruction of Ancestral Gene Orders*, in "ISCB-Latin America", Belo Horizonte, Brazil, October 2014, https://hal.inria.fr/hal-01083841.

[8] A. SAFFARIAN, M. GIRAUD, H. TOUZET. *Searching for alternate RNA structures in genomic sequences*, in "CMSR - Computational Methods for Structural RNAs", Strasbourg, France, 1st workshop on Computational Methods for Structural RNAs, September 2014, vol. 1, pp. 13-24 [*DOI :* 10.15455/CMSR.2014.0002], https://hal.archives-ouvertes.fr/hal-01084319

[9] C. VROLAND, M. SALSON, H. TOUZET. *Lossless seeds for searching short patterns with high error rates*, in "International Workshop On Combinatorial Algorithms", Duluth, United States, October 2014, https://hal.archives-ouvertes.fr/hal-01079840

National Conferences with Proceedings

[10] N. GRARDEL, M. SALSON, A. CAILLAULT, M. DUEZ, C. VILLENET, S. QUIEF, S. SEBDA, C. ROUMIER, M. FIGEAC, C. PREUDHOMME, M. GIRAUD. *Diagnostic et suivi des leucémies aiguës lymphoblastiques (LAL) par séquençage haut-débit (HTS)*, in "Congrès de la Société Française d'Hématologie (SFH)", Paris, France, 2014, https://hal.archives-ouvertes.fr/hal-01100152

Conferences without Proceedings

[11] E. KOPYLOVA, L. NOÉ, P. PERICARD, M. SALSON, H. TOUZET. *SortMeRNA 2: ribosomal RNA classification for taxonomic assignation*, in "ECCB - Workshop on Recent Computational Advances in Metagenomics", Strasbourg, France, September 2014, https://hal.archives-ouvertes.fr/hal-01094011

[12] E. KOPYLOVA, L. NOÉ, P. PERICARD, H. TOUZET. *SortMeRNA 2: ribosomal RNA classification for taxonomic assignation* , in "Bioinformatics for Environmental Genomics", Lyon, France, May 2014, https://hal.archives-ouvertes.fr/hal-01094029

Other Publications

[13] N. GRARDEL, M. GIRAUD, M. SALSON, M. DUEZ. *Multiclonal Diagnosis and MRD Follow-up in ALL with HTS Coupled with a Bioinformatic Analysis*, 2014, American Society of Hematology Annual Meeting, https://hal.archives-ouvertes.fr/hal-01100290

References in notes

[14] F. LIPMANN, W. GEVERS, H. KLEINKAUF, R. J. ROSKOSKI. *Polypeptide synthesis on protein templates: the enzymatic synthesis of gramicidin S and tyrocidine*, in "Adv Enzymol Relat Areas Mol Biol", 1971, vol. 35, pp. 1–34