IN PARTNERSHIP WITH:
CNRS

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2013

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]
Mathieu Giraud [CNRS, Researcher]
Aïda Ouangraoua [Inria, Researcher, on parental leave from January to February 2013]

Faculty Members

Stéphane Janot [Univ. Lille I, Associate Professor]
Laurent Noé [Univ. Lille I, Associate Professor]
Maude Pupin [Univ. Lille I, Associate Professor]
Mikaël Salson [Univ. Lille I, Associate Professor]
Jean-Stéphane Varré [Univ. Lille I, Professor, HdR]

Engineers

Thierry Barthel [Inria, from Oct 2012]
Jean-Frédéric Berthelot [CNRS, from Oct 2012]
Marc Duez [Univ. Lille I, from Mar 2013 until Dec 2013]
Areski Flissi [CNRS, from Jul 2013]
Alan Lahure [CNRS, from Jul 2013]

PhD Students

Yoann Dufresne [Univ. Lille I, from Oct 2013]
Evguenia Kopylova [Univ. Lille I, until Dec 2013]
Pierre Pericard [Univ. Lille I, from Nov 2013]
Antoine Thomas [Univ. Lille I, from Sep 2010]
Christophe Vroland [Univ. Lille I, from Oct 2012]

Post-Doctoral Fellow

Ammar Hasan [Inria, until Mar 2013]

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.lifl.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 [13], [20], [22], [16], [15]. Members of the team have also a strong expertise in text indexing and compressed index data structures [21], [24], [23]. 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 [25], [19], [18], [17], [11] or non-ribosomal peptides [12]. 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. It covers several paradigms: grids, single-instruction, multiple-data (SIMD) instructions or manycore processors such as graphics cards (GPU). For example, libraries like CUDA and OpenCL also facilitate the use of these manycore processors. These hardware architectures bring promising opportunities for time-consuming bottlenecks arising in bioinformatics.

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 their MCMC sampling allow to approximate probability distributions over parameters and to describe more 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 experienced remarkable advances with 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 allowing to gene expression analysis. 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 noncodingRNA analysis. Noncoding RNA genes play 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 acids 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 mathematical models 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. Software and Platforms

5.1. YASS – Local homology search

Actively maintained.
Software self-assessment following the mechanisms provided by Inria Evaluation Committee for software evaluation: **A-4**, **SO-3**, **SM-2**, **EM-3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-4

Software web site : http://bioinfo.lifl.fr/yass/

Licence: GPL

YASS is a software devoted to the classical problem of genomic pairwise alignment, and use most of our knowledge to design and implement efficient seeding techniques these last years. It is frequently used, it always receives more than 300 web queries per month (excluding local queries), and is also frequently downloaded and cited.

5.2. RNA tools – RNA structure prediction and comparison

http://bioinfo.lifl.fr/rna/

Actively maintained/Actively developed

Inria Evaluation Committee Criteria for Software Self-Assessment: A-4, SO-3, SM-2, EM-3, SDL-4, DA-4, CD-4, MS-4, TPM-4

The RNA tools provide a suite of programs to help analysing RNA secondary structures, together with visualisation tools for RNA 2D structures and RNA multiple alignments. Our first tool was *carnac* for RNA structure prediction by comparative analysis. *carnac* was issued in 2004 [1], independently benchmarked [2], and re-designed in 2009. It is still cited and used. Over the years, we have add new programs: *regliss* for locally optimal secondary structures, *gardenia* for structure comparison, *CG-seq* for gene prediction by comparative analysis, ...

5.3. TFM-Explorer – Identification and analysis of transcription factor binding sites

Actively maintained.
Software self-assessment: **A-4**, **SO-3**, **SM-2**, **EM-3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-4

Software web site : http://bioinfo.lifl.fr/TFM/

Licence: GPL

The TFM suite is a set of tools for analysis of transcription factor binding sites modeled by Position Weight Matrices. In this suite, the TFM-EXPLORER tool is designed to analyze regulatory regions of eukaryotic genomes using comparative genomics and local over-representation.

5.4. RNAspace – A platform for noncoding RNA annotation

Actively developped.
Software self-assessment: **A-5**, **SO-3**, **SM-3-up4**, **EM-2-up3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-4

Software web site : http://www.rnaspace.org/

Licence: GPL

[1] CARNAC: folding families of related RNAs. H. Touzet *et al.*, Nucleic Acids Research, 2004
[2] A comprehensive comparison of comparative RNA structure prediction approaches. P. Gardner *et al.*, BMC Bioinformatics, 2004

RNAspace is a national collaborative initiative conducted with Genopole Midi-Pyrénées and originally supported by IBISA [3]. The goal is to develop an open source platform for structural and functional noncoding RNA annotation in genomes (see Section 6.2): http://www.rnaspace.org. The project will be pursued within France Génomique (see Section 7.2.1).

5.5. CGseq – A toolbox for comparative analysis

Actively maintained.
Software self-assessment: **A-4**, **SO-3**, **SM-2**, **EM-3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-4

Software web site : http://bioinfo.lifl.fr/CGseq/

Licence: GPL

CG-seq is a toolbox for identifying functional regions in a genomic sequence by comparative analysis using multispecies comparison.

5.6. SortMeRNA – Metatranscriptome classification

Actively developed.
Software self-assessment: **A-4**, **SO-3**, **SM-2**, **EM-3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-4

Software web site: http://bioinfo.lifl.fr/RNA/sortmerna

Licence: GPL

\+ *SortMeRNA*: Metatranscriptome classification

http://bioinfo.lifl.fr/RNA/sortmerna/

Actively developed

Inria Evaluation Committee Criteria for Software Self-Assessment: A-4, SO-3, SM-3, EM-3, SDL-4, DA-4, CD-4, MS-4, TPM-4

SortMeRNA is a tool designed to rapidly filter ribosomal RNA fragments from metatranscriptomic data produced by next-generation sequencers. The distribution includes curated ribosomal RNA databases. It is available for download from our website, or through the open web-based platform Galaxy. *SortMeRNA* was released in October 2012, and is used in production by Genoscope (French National Center for Sequencing) to process metatranscriptomic data. Moreover, it has already been integrated in two published computational pipelines [4, 5] and have identified users in multiple research laboratories worldwide [6].

5.7. Vidjil – Quantifying lymphocytes rearrangements in high-throughput sequencing data

Actively developed
Software self-assessment: **A-3-up4**, **SO-3**, **SM-2-up3**, **EM-3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-4

Software web site : http://bioinfo.lifl.fr/vidjil

[3] IBISA is a French consortium for evaluating and funding national technological platforms in life sciences.

[4] A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. M. Leimena et al., *BMC genomics*, 2013

[5] Metagenome survey of a multispecies and algae-associated biofilm reveals key elements of bacterial-algae interactions in photobioreactors. I. Krohn-Molt et al. Applied and environmental microbiology, 2013

[6] Umeå University (Sweden), Leibniz Institute DSMZ (Germany), NGS department of Campus Science Support Facilities GmbH (Austria), Oxford Centre for Integrative Systems Biology (Great Britain), Laboratoire d'Ecologie Alpine (Grenoble), PRABI (Lyon), Wageningen University (Netherlands), ...

Vidjil implements a two-stage strategy for fast clustering and quantification of clones coming from immunological rearrangements in genomic sequences. It is currently used in "minimal residual disease" following, but could have other uses in immunology research. *Vidjil* is currently under test at the Lille hospital, and is planned to be tested in another hematological lab. In 2013, the development of Vidjil was supported by the regional project ABILES: an engineer (Marc Duez) developed for 5 months a graphical interface for using *Vidjil*. We plan to release a first production version to the hospital during 2014.

5.8. Biomanycores.org – A community for bioinformatics on manycore processors

Actively developped.
Software self-assessment: **A-3**, **SO-2**, **SM-3**, **EM-3down2**, **SDL-4up5**, OC-4 (DA-4, CD-4, MS-4, TPM-4)

Software web site : http://biomanycores.org/

Manycore architectures are an emerging field of research full of promises for parallel bioinformatics. However the usage of GPUs is not so widespread in the end-user bioinformatics community. The goal of the `biomanycores.org` project is to gather open-source CUDA and OpenCL parallel codes and to provide easy installation, benchmarking, and interoperability. The last point includes interfaces to popular frameworks such as Biopython, BioPerl and BioJava.

The development of Biomanycores was supported by a national ADT [7] from October 2010 to October 2012.

5.9. Norine – A resource for nonribosomal peptides

Actively maintained.
Software self-assessment: **A-5**, **SO-3**, **SM-3-up4**, **EM-2-up3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-4

Software web site : http://bioinfo.lifl.fr/norine/ 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 nonribosomal 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. Norine main users come for 13% from the United States of America, for 12% from the United Kingdom, for 5% from China or for 4% from Germany where renowned biology laboratories work on nonribosomal peptides (NRPs) or on their synthetases.

5.10. Crac – RNA-seq read analysis

Actively maintained.
Software self-assessment: **A-4**, **SO-3**, **SM-3**, **EM-3**, **SDL-4**, DA-4, CD-4, MS-4, TPM-3

Software web site: http://crac.gforge.inria.fr/

Objective: CRAC aims at identifying biological variations in RNAs by comparing short reads to a reference genome. It detects point mutations, short indels, splice events, and fusion genes or transcripts.

This library is the result of a collaboration with N. Philippe and T. Commes (IGH laboratory, Montpellier) and É. Rivals (LIRMM laboratory, Montpellier).

5.11. GkArrays – Indexing high throughput sequencer reads

Actively maintained.
Software self-assessment: **A-3**, **SO-3**, **SM-3**, **EM-2**, **SDL-4**, DA-4, CD-4, MS-4, TPM-3

[7] ADT (Action for Technological Development) is an Inria internal call

Software web site : http://crac.gforge.inria.fr/gkarrays/

Objective : Gk-Arrays is a C++ library specifically dedicated to indexing reads produced by high-throughput sequencers. This index allows to answer queries centred on reads. It also takes benefits from the input specificity to lower space consumption.

This library is the result of a collaboration with N. Philippe and T. Commes (IGH laboratory, Montpellier), M. Léonard and T. Lecroq (LITIS laboratory, Rouen) and É. Rivals (LIRMM laboratory, Montpellier).

6. New Results

6.1. High-throughtput sequence processing

- Within our collaboration with Montpellier (IRB and LIRMM) we published a paper on CRAC, a software for analysing short RNA sequences and detecting variations among them [5].
- We have been invited to contribute an invited book chapter on metatranscriptomic data analysis (*Methods in Molecular Biology*, in press). This chapter covers the complete bioinformatic analysis from raw reads to taxonomic assignation, and introduces our software SortMeRNA (see Paragraph 5.6). This is a joint work with team LABIS in Genoscope.
- Evguenia Kopylova defended her thesis on December, the 11th (*"New algorithmic and bioinformatic approaches for the analysis of data from high throughput sequencing"*, [1]). The second part of her work deals with a new read mapper for metagenomic sequence data.
- Within our collaboration with the Lille hospital, we developped a seed-based heuristics for the detection of lymphocyte rearrangements from high-throughput data. This method is implemented in the software Vidjil (see Section 5.7). Our results were presented at the Jobim conference [8], and a journal article was submitted.

6.2. RNA algorithms

- We have started a new collabarative project with Bielefeld Universität on an extension of *Algebraic Dynamic Programming*. We introduced a generic specification framework, called *inverted coupled rewrite systems* [9], that can deal with optimization problems on strings, trees, and arc-annotated sequences. 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.

6.3. Genomic rearrangements

- Within a collaboration with LIAFA (CNRS UMR 7089, and University Paris 7) we published a method for the assembling of ancestral gene orders from contiguous ancestral fragments [4].

6.4. Nonribosomal peptides

- Yoann Dufresne is starting a PhD thesis on computational biology for nonribosomal peptides (NRPs) under the supervision of Maude Pupin and Laurent Noe, after doing his master thesis with them. He already worked on the translation of the chemical structure of the NRPs into their monomeric structure. NRPs can be represented by their chemical structure that is a graph where the atoms are represented by nodes and the chemical bonds by arcs; or by their monomeric structure that is a graph where the monomers are represented by nodes and the chemical bonds between monomers by arcs. We designed a novel algorithm capable of localizing the monomers from a reference list in the chemical structures of peptides [7]. It is based on a heuristic that utilizes chemical information of NRPs. The preliminary results are encouraging, and should lead to further studies.

7. Partnerships and Cooperations

7.1. 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.
- Projet émergent call 2011. "ABILES – Algorithmes bioinformatiques pour le diagnostic de leucémie résiduelle par séquenceurs haut-débit" involving IRCL (Institut de recherche sur le cancer de Lille, Inserm, Université Lille 2), Hematology department of Lille Hospital and BONSAI (see the Vidjil software, Section 5.7).

7.2. National Initiatives

7.2.1. ANR

- ANR Mappi (2010-2013): National funding from the French Agency Research (call *Conception and Simulation*). This project involves four partners: LIAFA (Université Paris 7), Genescale (Inria Rennes), Genoscope (French National Center for Sequencing) and BONSAI. The topic is *Nouvelles approches algorithmiques et bioinformatiques pour l'analyse des grandes masses de données issues des séquenceurs de nouvelle génération*.
- 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 (2012): 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 *Silenes* (2012-2013): National funding from CNRS. This project involves the GEPV (P. Touzet) and the IBMP [8] (J. Gualberto, L. Maréchal-Drouard). The topic is *Etude comparative de l'architecture du génome mitochondrial chez les Caryophyllacées et les Poacées*. It aims to sequence and analyze the genome structure of a number of *Silene* ecotypes and to compare them to other species.
- 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. We are two partners: UMR 8199 (Génomique et maladie métabolique, Ph Froguel, O. Sand, part of the LIGAN sequencing platform) and BONSAI.

7.2.2. PEPS

- PEPS Biology-Mathematics-Computer science: "Etude comparative de l'architecture du génome mitochondrial chez les Caryophyllacées et les Poacées". This project involves three partners: IBMP (Institut de Biologie Moléculaire des Plantes), GEPV (UMR CNRS 8198, Université Lille 1) and BONSAI.
- PEPS Biology-Mathematics-Computer science: "Algorithmes pour l'alignement des lectures et la découverte de variants dans les projets de reséquençage". This project involves two partners: UMR 8199 Génomique et Maladies Métaboliques and BONSAI.

[8] Institut de Biologie Moléculaire des Plantes - UPR2357, Strasbourg

7.2.3. ADT

- ADT biosciences resources (2011-2013): 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. An engineer was hired from 2011 to 2013 and worked in Rennes and another one was hired in 2012 and works in Lille.

7.3. International Initiatives

7.3.1. Inria International Partners

7.3.1.1. Informal International Partners

- *Universität Tübingen :* We have a collaboration with Tilmann Weber on the topic of computational biology for nonribosomal peptides. We co-organized a workshop in Lille with him.
- We have a collaboration with Martin C. Frith from the *Computational Biology Research Center* (Tokyo) on the topic of transition spaced seeds.
- *LaCIM (Laboratoire de Combinatoire et d'Informatique Mathématique):* Since 2009, we have been collaborating with Anne Bergeron (Univ. du Québec à Montréal), Krister Swenson (Univ. de Montréal), and Cédric Chauve (Simon Fraser Univ.) on theroretical and applied aspects of gene orders evolution. In 2011, we began a new project on the analysis of exonic gene structure evolution.
- *Universität Bielefeld (Germany)*: This collaboration started through a PHC Procope bilateral cooperation project with the team of Pr. Robert Giegerich (2010-2011). The goal was to work on a generic parallelization of the Algebraic Dynamic Programming methodology. This partnership is still ongoing, with several visits of Robert Giegerich these last few months. It is the source of our recent work for an extension of Algebraic Dynamic Programming [9].

7.4. International Research Visitors

7.4.1. Visits of International Scientists

The following scientists visited the team and gave a talk at the team or the laboratory seminar:

- Mihai Pop, University of Maryland (28 may)
- Veli Mäkinen, university of Helsinki (11 december)
- Krister M. Swenson, UQAM (12 november)

8. Dissemination

8.1. Scientific Animation

- The team actively participates in the national GDR *Bioinformatique moléculaire*. H. Touzet has been a member of the executive committee since 2007. In this context, she coorganized a two-day workshop, called Seqbio, in Montpellier in November 2013
- We organize a regular pluridisciplinary seminar on bioinformatics, whose audience is composed of researchers in biology and bioinformatics. In the last twelve months, we proposed three events: *Metagenomics* (110 participants), *Phylogenomics* (52 participants) and *Structural bioinformatics* (30 participants)
- We organized in Lille, with our collaborator Tilmann Weber from Universität Tüebingen, an international workshop on "Bioinformatics tools for NRPS discovery" in July. The schedule included introducing lectures by invited speakers which are key scientists in the field and practical sessions. It gathered 30 scientists from all the continents.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Our research work finds also its expression in a strong commitment in pedagogical activities at the University Lille 1. For several years, members of the project have been playing a leading role in the development and the promotion of bioinformatics (more than 400 teaching hours per year). We are involved in several graduate diplomas (research master degree) in computer science and biology (*master biologie-santé, master génomique et protéomique, master biologie-biotechnologie*) in an Engineering School (Polytech'Lille), as well as in permanent education (for researchers, engineers and technicians).

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

M. Salson, *Coding and information theory*, 36h, L2 (licence "Computer science", univ. Lille 1)

J.-S. Varré *Programming with Caml*, 55h, L2 (licence "Sciences for Engineers", univ. Lille 1)

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

L. Noé, *Networks*, 36h, L3 (licence "Computer science", univ. Lille 1)

L. Noé, *System*, 36h, L3 (licence "Computer science", univ. Lille 1)

M. Pupin, *Databases*, 36h, L3 (licence "Computer science", univ. Lille 1)

M. Pupin, *Professional project*, 18h, L3 (licence "Computer science", univ. Lille 1)

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

S. Janot, *Introduction to programming*, 50h, first year of engineering school (L3) (Polytech'Lille, univ. Lille1)

S. Janot, *Introduction to databases*, 30h, first year of engineering school (L3) (Polytech'Lille, univ. Lille1)

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

L. Noé, *Individual project*, organiser, M1 (master "Computer science", univ. Lille 1)

M. Pupin, *Introduction to programming (JAVA)*, 30h, M1 (master "Mathématiques et finance", univ. Lille 1)

M. Salson, J.-S. Varré, *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)

S. Janot, *Databases*, 12h, second year of engineering school (M1) (Polytech'Lille, univ. Lille1)

S. Janot, *Introduction to artificial intelligence*, 25h, second year of engineering school (M1) (Polytech'Lille, univ. Lille1)

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

M. Pupin, *Practical bioinformatics*, 35h, M2 (master "Génomique Protéomique", univ. Lille 1)

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

J.-S. Varré, *ISN - Computer science for secondary school*, 30h, second-level teachers.

8.2.2. Supervision

- HDR: *Maude Pupin*, Modèles bio-informatiques pour les peptides non-ribosomiques et leurs synthétases. Defense in December 2013 [2].
- PhD : *Evguenia Kopylova*, New algorithmic and bioinformatic approaches for the analysis of data from next-generation sequencing, Université Lille 1, co-directed by H. Touzet and L. Noé. Thesis defended in December 2013 [1].
- PhD in progress : *Christophe Vroland*, microRNA repertoire and target evolution: developing efficient indexing techniques and comparison between close plant species, Université Lille 1, co-directed by H. Touzet, M. Salson from BONSAI and V. Castric ("Genetics and evolution in plants" laboratory).
- PhD in progress : *Pierre Péricard*, high-throughput sequencing : taxonomic assignation of metaomic sample reads, Université Lille 1, co-directed by H. Touzet and S. Blanquart.
- PhD in progress : *Yoann Dufresne*, Models and algorithms to analyse and predict non-ribosomal peptides, Université Lille 1, co-directed by M. Pupin and L. Noé.

8.2.3. Juries

- Member of the thesis committee of Natalia Golenetskaya (Université Bordeaux 1, J.-S. Varré)
- Member of the habilitation committee of F. Jossinet (Université de Strasbourg, H. Touzet) and C. Lhoussaine (Université Lille 1, H. Touzet)

8.2.4. Administrative activities

- National representative (*chargée de mission*) for the Institute for Computer Sciences (INS2I) in CNRS [9]. 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 (H. Touzet)
- Member of the Gilles Kahn PhD award commitee (H. Touzet)
- Member of ITMO Genetics, Genomics and Bioinformatics of AVIESAN (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)
- 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)
- Head of the GIS department (Statistics and Computer Sciences) of Polytech'Lille (S. Janot)
- Member of the LIFL Laboratory council (L. Noé, H. Touzet)
- We made two public science presentations (leukemia and high-throughput sequecing, M. Duez, M. Giraud, M. Salson, and transcript comparisons, A. Ouangraoua)
- This year, we did not have a significant activity in high schools due to schedule constraints during the "week of science". We plan to relaunch this activity in 2014.

[9] CNRS: National Center for Scientific Research

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] E. KOPYLOVA. , *Algorithmes bio-informatiques pour l'analyse de données de séquençage à haut débit*, Université des Sciences et Technologie de Lille - Lille I, December 2013, http://hal.inria.fr/tel-00919185

[2] M. PUPIN. , *Modèles bio-informatiques pour les peptides non-ribosomiques et leurs synthétases*, Université des Sciences et Technologie de Lille - Lille I, December 2013, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00918918

Articles in International Peer-Reviewed Journals

[3] M. GROUSSIN, B. BOUSSAU, S. CHARLES, S. BLANQUART, M. GOUY. *The molecular signal for the adaptation to cold temperature during early life on Earth*, in "Biol Lett", October 2013, vol. 9, n^{o} 5, http://hal.inria.fr/hal-00918283

[4] A. OUANGRAOUA, M. RAFFINOT. *On the Identification of Conflicting Contiguities in Ancestral Genome Reconstruction*, in "Journal of Computational Biology", 2013, pp. 1-16 [*DOI :* 10.1089/CMB.2013.0086], http://hal.inria.fr/hal-00913212

[5] N. PHILIPPE, M. SALSON, T. COMMES, E. RIVALS. *CRAC: an integrated approach to the analysis of RNA-seq reads*, in "Genome Biology", March 2013, vol. 14, n^{o} 3 [*DOI :* 10.1186/GB-2013-14-3-R30], http://hal.inria.fr/inserm-00850972

[6] B. REISINGER, J. SPERL, A. HOLINSKI, V. SCHMID, C. RAJENDRAN, L. CARSTENSEN, S. SCHLEE, S. BLANQUART, R. MERKL, R. STERNER. *Evidence for the Existence of Elaborate Enzyme Complexes in the Paleoarchean Era*, in "Journal of the American Chemical Society", December 2013 [*DOI :* 10.1021/JA4115677], http://hal.inria.fr/hal-00924047

National Conferences with Proceedings

[7] Y. DUFRESNE, V. LECLÈRE, P. JACQUES, L. NOÉ, M. PUPIN. *Non Ribosomal Peptides : A monomeric puzzle*, in "JOBIM - Journées Ouvertes en Biologie, Informatique et Mathématiques", Toulouse, France, July 2013, pp. 143-150, http://hal.inria.fr/hal-00843827

[8] M. GIRAUD, M. SALSON, M. DUEZ, J.-S. VARRÉ, C. VILLENET, S. QUIEF, A. CAILLAULT, N. GRARDEL, C. ROUMIER, C. PREUDHOMME, M. FIGEAC. *Suivi de la leucémie résiduelle par séquençage haut-débit*, in "JOBIM", Toulouse, France, July 2013, http://hal.inria.fr/hal-00857581

Conferences without Proceedings

[9] R. GIEGERICH, H. TOUZET. *Algebraic Dynamic Programming 2.0*, in "Workshop Haskell-Treffen an der Universität Leipzig", Leipzig, Germany, June 2013, http://hal.inria.fr/hal-00857801

Research Reports

[10] A. ROUSSEAU, A. DARNAUD, B. GOGLIN, C. ACHARIAN, C. LEININGER, C. GODIN, C. HOLIK, C. KIRCHNER, D. RIVES, E. DARQUIE, E. KERRIEN, F. NEYRET, F. MASSEGLIA, F. DUFOUR, G. BERRY, G. DOWEK, H. ROBAK, H. XYPAS, I. ILLINA, I. GNAEDIG, J. JONGWANE, J. EHREL, L. VIENNOT, L. GUION, L. CALDERAN, L. KOVACIC, M. COLLIN, M.-A. ENARD, M.-H. COMTE, M. QUINSON, M. OLIVI, M. GIRAUD, M. DORÉMUS, M. OGOUCHI, M. DROIN, N. LACAUX, N. ROUGIER, N. ROUSSEL, P. GUITTON, P. PETERLONGO, R.-M. CORNUS, S. VANDERMEERSCH, S. MAHEO, S. LEFEBVRE, S. BOLDO, T. VIÉVILLE, V. POIREL, A. CHABREUIL, A. FISCHER, C. FARGE, C. VADEL, I. ASTIC, J.-P. DUMONT, L. FÉJOZ, P. RAMBERT, P. PARADINAS, S. DE QUATREBARBES, S. LAURENT. , *Médiation Scientifique : une facette de nos métiers de la recherche*, March 2013, 34 p. , http://hal.inria.fr/hal-00804915

References in notes

[11] G. BLIN, A. DENISE, S. DULUCQ, C. HERRBACH, H. TOUZET. *Alignment of RNA structures*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2008, http://dx.doi.org/10.1109/TCBB.2008.28

[12] S. CABOCHE, M. PUPIN, V. LECLÈRE, P. JACQUES, G. KUCHEROV. *Structural pattern matching of nonribosomal peptides*, in "BMC Structural Biology", March 18 2009, vol. 9:15 [*DOI :* 10.1186/1472-6807-9-15], http://www.biomedcentral.com/1472-6807/9/15

[13] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *Subset Seed Automaton*, in "12th International Conference on Implementation and Application of Automata (CIAA 07)", Lecture Notes in Computer Science, Springer Verlag, 2007, vol. 4783, pp. 180–191 [*DOI :* 10.1007/978-3-540-76336-9_18], http://www.springerlink.com/content/y824l20554002756/

[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

[15] L. NOÉ, M. GÎRDEA, G. KUCHEROV. *Designing efficient spaced seeds for SOLiD read mapping*, in "Advances in Bioinformatics", July 2010, vol. 2010 [*DOI :* 10.1155/2010/708501], http://www.hindawi.com/journals/abi/2010/708501/

[16] L. NOÉ, M. GÎRDEA, G. KUCHEROV. *Seed design framework for mapping SOLiD reads*, in "Proceedings of the 14th Annual International Conference on Research in Computational Molecular Biology (RECOMB), April 25-28, 2010, Lisbon (Portugal)", B. BERGER (editor), Lecture Notes in Computer Science, Springer, April 2010, vol. 6044, pp. 384–396 [*DOI :* 10.1007/978-3-642-12683-3_25], http://www.springerlink.com/content/41535x341gu34131/

[17] A. OUANGRAOUA, P. FERRARO. *A constrained edit distance algorithm between semi-ordered trees*, in "Theor. Comput. Sci.", 2009, vol. 410, n^{o} 8-10, pp. 837-846

[18] A. OUANGRAOUA, P. FERRARO. *A new constrained edit distance between quotiented ordered trees*, in "J. Discrete Algorithms", 2009, vol. 7, n^{o} 1, pp. 78-89

[19] A. OUANGRAOUA, P. FERRARO, L. TICHIT, S. DULUCQ. *Local similarity between quotiented ordered trees*, in "J. Discrete Algorithms", 2007, vol. 5, n^{o} 1, pp. 23-35

[20] P. PETERLONGO, L. NOÉ, D. LAVENIER, G. GEORGES, J. JACQUES, G. KUCHEROV, M. GIRAUD. *Protein similarity search with subset seeds on a dedicated reconfigur able hardware*, in "Parallel Processing and

Applied Mathematics / Parallel Biocomputi ng Conference (PPAM / PBC 07)", R. WYRZYKOWSKI, J. DONGARRA, K. KARCZEWSKI, J. WASNIEWSKI (editors), Lecture Notes in Computer Science (LNCS), 2008, vol. 4967, pp. 1240-1248 [*DOI :* 10.1007/978-3-540-68111-3], http://www.lifl.fr/~giraud/publis/peterlongo-pbc-07.pdf

[21] P. PETERLONGO, L. NOÉ, D. LAVENIER, V. H. NGUYEN, G. KUCHEROV, M. GIRAUD. *Optimal neighborhood indexing for protein similarity search*, in "BMC Bioinformatics", 2008, vol. 9, n^{o} 534 [*DOI :* 10.1186/1471-2105-9-534], http://www.biomedcentral.com/1471-2105/9/534

[22] M. ROYTBERG, A. GAMBIN, L. NOÉ, S. LASOTA, E. FURLETOVA, E. SZCZUREK, G. KUCHEROV. *On subset seeds for protein alignment*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2009, vol. 6, n^{o} 3, pp. 483–494, http://www.lifl.fr/~noe/files/pp_TCBB09_preprint.pdf

[23] M. SALSON, T. LECROQ, M. LÉONARD, L. MOUCHARD. *A Four-Stage Algorithm for Updating a Burrows-Wheeler Transform*, in "Theoretical Computer Science", 2009, vol. 410, n^{o} 43, pp. 4350–4359

[24] M. SALSON, T. LECROQ, M. LÉONARD, L. MOUCHARD. *Dynamic Extended Suffix Array*, in "Journal of Discrete Algorithms", 2010, vol. 8, pp. 241–257

[25] H. TOUZET. *Comparing similar ordered trees in linear-time*, in "Journal of Discrete Algorithms", 2007, vol. 5, n^{o} 4, pp. 696-705 [*DOI :* 10.1016/J.JDA.2006.07.002], http://linkinghub.elsevier.com/retrieve/pii/S1570866706000700