IN PARTNERSHIP WITH:
**CNRS**

**Université des sciences et technologies de Lille (Lille 1)**

# Activity Report 2012

# 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 and Bioinformatics**

# Table of contents

## Project-Team BONSAI

**Keywords:** Computational Biology, Genomics, Genome Annotation, High Performance Computing, Non-coding RNA, Nonribosomal Peptides, Genome Rearrangement

*Creation of the Project-Team:* January 01, 2011 .

# 1. Members

**Research Scientists**
Hélène Touzet [Team leader, Senior Researcher CNRS, HdR]
Samuel Blanquart [Junior Researcher Inria]
Mathieu Giraud [Junior Researcher CNRS]
Aïda Ouangraoua [Junior Researcher Inria, on maternal leave from April to July 2012]

**Faculty Members**
Stéphane Janot [Associate Professor, Université Lille 1]
Laurent Noé [Associate Professor, Université Lille 1]
Maude Pupin [Associate Professor, Université Lille 1]
Mikaël Salson [Associate Professor, Université Lille 1]
Jean-Stéphane Varré [Professor, Université Lille 1, HdR]

**External Collaborator**
Mohcen Benmounah [software engineer, Université Lille 1 since April 2012]

**Engineers**
Thierry Barthel [IJD, Inria, from October 2012]
Jean-Frédéric Berthelot [IJD, Inria, until October 2012]

**PhD Students**
Tuan Tu Tran [Inria CORDI fellowship, from September 2009]
Antoine Thomas [MESR fellowship, from October 2010]
Evguenia Kopylova [ANR grant, from December 2010]
Christophe Vroland [MESR fellowship, from October 2012]

**Post-Doctoral Fellow**
Ammar Hasan [Inria PostDoc grant, from February 2012]

**Administrative Assistants**
Sandrine Catillon [Inria, until October 2012]
Amélie Supervielle [Inria, from October 2012]

# 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 BONSAIṀost 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. Scientific Foundations

## 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 [20], [27], [29], [23], [22]. Members of the team have also a strong expertise in text indexing and compressed index data structures [28], [31], [30]. 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 [32], [26], [25], [24], [17] or non-ribosomal peptides [18]. 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.

## 3.2. High-performance computing

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.

## 3.3. 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 contexte, 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 allows 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 onto linkage maps and mathematical models appeared fifteen years ago. But the usage of computational tools was still limited because of 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 year 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 [21]. 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

## 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. Carnac – RNA structure prediction

*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/carnac/

Licence: Cecill

The CARNAC program is for RNA structure prediction by comprative analysis. The web interface also offers 2D visualisation tools and alignment functionalities with Gardenia. It has proven to be very fast and very specific compared to its competitors [19].

## 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. Regliss – RNA locally optimal structures

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

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

REGLISS is a tool that studies the energy landscape of a given RNA sequence by generating all locally optimal structures, that are maximal thermodynamically stable structures.

## 5.5. 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

RNAspace is a national collaborative initiative conducted with Genopole Midi-Pyrénées and originally supported by IBISA [1]. 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).

## 5.6. 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.

## 5.7. 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 is a software designed to rapidly filter ribosomal RNA fragments from metatransriptomic data produced by next-generation sequencers. It is capable of handling large RNA databases and sorting out all fragments matching to the database with high accuracy and specificity.

## 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 [2] between BONSAI, SYMBIOSE (CRI Rennes) and DOLPHIN (CRI Lille), from October 2010 to October 2012. This ADT led to the hiring of J.-F. Berthelot (IJD) who completely redesigned the existing code and added more applications. Biomanycores has now a comprehensive developer and user documentation, large test suites and continuous integration. In June 2012, the project has been presented during a workshop dedicated to manycore programming (see Section 8.1).

## 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

[1] IBISA is a French consortium for evaluating and funding national technological platforms in life sciences.
[2] ADT (Action for Technological Development) is an Inria internal call

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 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. 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. 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

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 the PhD of T. T. Tran, we proposed a new indexing structure adapted to GPUs. We studied an indexing scheme with perfect hashing functions, and developed a prototype written in openCL for a read mapper. This read mapper has a sensitivity comparable to state-of-the-art read mappers, and provides substantial time gains in some cases.
- Within our collaboration with the Lille hospital on the follow-up of leukemia residual disease, we proposed a new heuristic to study immunological VDJ recombinations and follow their evolution along the time. The method is under testing on several datasets obtained from the Ion Torrent sequencer at IRCL (Institut de Recherche sur le Cancer de Lille).
- Within the PhD of E. Kopylova, we designed an new algorithm to filter out ribosomal RNA sequences from RNA raw data produced in metatranscriptomic sequencing. The method combines text indexing techniques, with the Burst trie, and Universal Levenshtein automaton to allow for seraching with errors. An article has been published the journal *Bioinformatics* [4].

## 6.2. Noncoding RNAs

- We designed a new algorithm to produce all locally optimal secondary structures of an RNA sequence. Locally optimal secondary structures are thermodynamically stable RNA structures that are maximal for inclusion: they cannot be extended without producing a conflict between base pairs in the secondary structure, or increasing the free energy. This was published in *Journal for Computational Biology* [7].
- We took part to a collaborative work on benchmarking for RNA structure comparison. This work has been published in *Advances in bioinformatics* [2].

## 6.3. Genomic rearrangements

- Within the context of the PhD of A. Thomas:
  - We designed an algorithm for finding the minimal number of block interchanges required to transform a duplicated linear genome into a tandem duplicated linear genome. We provide a formula for the distance as well as a polynomial time algorithm for the sorting problem. This work was published in the conference *Bioinformatics* [13].
  - We explored a new problem concerning tandem halving, that is reconstructing a non-duplicated ancestor to a partially duplicated genome in a model where duplicated content is caused by several tandem duplications. We provide a distance in $O(n)$ time and a scenario in $O(n^2)$ time. We considered several problems related to multiple tandem reconstruction and proved that the simpliest of reconstructing 2 tandems is NP-hard. This work was published in the conference WABI 2012 [14].
- In the context of ancestral genome reconstruction, we designed an algorithm for the identification of Minimal Conflicting Sets (MCS) rows in a biological binary matrix. We provided a $O(n^2m^2 + nm^7)$ time algorithm, largely improving the up-to-date best algorithm in $O(m^6n^5(m+n)^2log(m+n))$ time. This work was published in the conference CPM 2012 [11].
- In the context of the comparison of sets of alternative gene transcripts, we designed a general framework to compare sets of transcripts that are transcribed from orthologous loci of several species. The model is based on the construction of a common reference sequence, and on annotations that allow the reconstruction of ancestral sequences, the identification of conserved events, and the inference of gains and losses of donor/acceptors sites, exons, introns and transcripts. This work was published in the conference ISBRA 2012 [12].

## 6.4. Nonribosomal peptides

- With the arrival of Ammar Hasan, a postdoc researcher, we started a new project on the prediction of nonribosomal peptides activity. We defined a novel peptide fingerprint based on monomer composition. This fingerprints is used for peptide similarity searching and for activity prediction. This work was published in *Journal of Computer-Aided Molecular Design* [1].
- We participated in the writing of a review dedicated to kurstakin, a nonribosomal lipopeptide synthetized by several Bacillus genus and published in *Applied microbiology and biotechnology* [3].
- The collaboration with members of EPI Orpailleur (CRI Nancy) succeeded in designing a protocol to discover new nonribosomal peptide synthetases in bacterial genomes and then annotate them in order to predict the peptide they produce. It was published in JOBIM 2012 [16].

# 7. Partnerships and Cooperations

## 7.1. Regional Initiatives

- At the end of 2010, we started a collaboration with the sequencing platform of Université Lille 2 and IRCL (M. Figeac) and the hematology lab of Lille hospital (N. Grardel, C. Roumier, C. Preudhomme), on the diagnosis of leukemia residual disease. This project has been awarded by a “Projet émergent region” grant for 2012 – 2013.
- Our research on *nonribosomal peptide synthesis* is based on a collaboration with the ProBioGEM laboratory (*Laboratoire des Procédés Biologiques Génie Enzymatique et Microbien*, Université Lille 1). This laboratory develops methods to produce and extract active peptides in agriculture or food. Two PhD thesis has been co-supervised by the two labs.
- We have a long term collaboration with GEPV Lab (Genetics and Evolution in Plants, UMR CNRS 8198, Université Lille 1). Topics includes rearrangements in mitochondrial genomes and evolution of plant miRNAs. One supervised PhD thesis has been defended in 2010, and a new thesis just started in October 2012.

- The team is in charge of the PPF *Bioinformatique*. This is an initiative of Université Lille 1 that coordinates public bioinformatics activities at the local level. It gathers seven labs coming from biology, biochemistry and computer science. Main topics are proteomics, microbiology, population genetics, etc.

## 7.2. National Initiatives

### *7.2.1. ANR*

- ANR Mappi (2010-2013, call *Conception and Simulation*). This project involves four partners: LIAFA (Université Paris 7), Genscale (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*.
- ANR France Génomique (2011-2014, PIA *Infrastructures Biologie Santé*). This national project involves 13 partners, including sequencing paltforms and bioinformatics platform. We take part to the workpackage on sRNA-seq data analysis.

### *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.

### *7.2.3. ADT*

- ADT biomanyocres (2010-2012): see section 5.8.
- 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.

## 7.3. International Initiatives

- S. Blanquart pursues his collaboration with the Sterner Group of the "Institut für Biophysik und Physikalische Biochemie" (Regensburg, Germany) on an ancestral sequences resurrection project. Researchers of the Sterner Groups succeeded in the resurrection and characterization of the LUCA's (Last Universal Common Ancestor) Histidine F enzyme, which have a TIM barrel fold. The paleo-enzyme works fine, just as do modern ones. It is the oldest resurrected yet proteins to our knowledge.
- In genomic rearrangement, we pursued our collaboration with the LaCIM at Université du Québec à Montréal, and DIRO at Université de Montréal. In the context of multiple genome comparison, we proposed a new framework for the multiple comparison of sets of transcripts transcribed from orthologous loci of several species [12].

## 7.4. International Research Visitors

### *7.4.1. Visits to International Teams*

- A. Thomas, Univ. du Québec à Montréal (Canada), visit to Anne Bergeron (2 weeks),
- J.-S. Varré, Univ. du Québec à Montréal (Canada), visit to Anne Bergeron (1 week),
- A. Ouangraoua, Univ. du Québec à Montréal (Canada), visit to Anne Bergeron (4 months),
- M. Giraud, Univ. of Thessaloniki (Greece), visit to E. Cambouroupoulos (1 month).

# 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, we hosted the national annual workshop, Seqbio, devoted to sequence analysis and bioinformatics (2 days, 80 participants, december 2011).

- 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: *manycores programming in biology* (2 days, 35 participants), *phylogeny* (2 days, 38 participants), *Analysis of NGS data* (1 day, 110 participants).
- We organized the annual meeting of the GTGC workgroup (Comparative Genomics Working Group) (1.5 days, 40 participants, december 2012).

## 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, *Algorithms (Ada)*, 58h, L3 (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)

A. Ouangraoua *Programming with MATLAB*, 36h, L1 (licence “Mechanical Engineering”, École de Technologie Supérieure de Montréal)

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)

M. Giraud, L. Noé, M. Pupin, *High-performance bioinformatics*, 28h, M2 (master “Calcul Scientifique”, univ. Lille 1)

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

### 8.2.2. Supervision

PhD : *Tuan Tu Tran*, Bioinformatics Sequence Comparisons on Manycore Processors, Univ. Lille 1, defense scheduled on 21 December 2012, co-directed by J.-S. Varré and M. Giraud

PhD : *Aurélien Vanvlassenbroeck*, Experimental and *in silico* study of the nonribosomal synthesis done by fluorescent *Pseudomonas*, Université Lille 1, 17 July 2012, co-directed by M. Pupin, V. Leclère (ProBioGEM lab) and P. Jacques (ProBioGEM lab)

PhD in progress : *Antoine Thomas*, Algorithms for genome rearrangement with duplications, Université Lille 1, co-directed by J-S. Varré and A. Ouangraoua.

PhD in progress : *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.

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).

### 8.2.3. Juries

- Member of the thesis committee of W. Abdelwahed (Univ. Lille 1, M. Pupin) and S. Benabderrahmane (Univ. Nancy 1, M. Pupin), Ph. Bordron, (Univ. Nantes, J-S. Varré, L. Noé), Benoit Groz (Univ. Lille 1, H. Touzet), Tarek El Falah (Univ. Rouen, J.-S. Varré)
- Member of the habilitation committee of J. Bourdon (Univ. Nantes, H. Touzet)

### 8.2.4. Administrative activities

- National representative (*chargée de mission*) for the Institute for Computer Sciences (INS2I) in CNRS [3]. 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)
- Scientific secretary of the Gilles Kahn PhD award commitee (M. Giraud)
- Member of ITMO Genetics, Genomics and Bioinformatics of AVIESAN (H. Touzet)
- Member of CSS MBIA (mathematics, bioinformatics and artificial intelligence) at INRA (H. Touzet)
- Head of PPF bioinformatics – University Lille 1 (H. Touzet)
- Head of Bilille, Lille bioinformatics platform (M. Pupin)
- Head of ReNaBi-NE (pôle Nord-Est du Réseaux National 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)
- Member of hiring committee (jury d'audition) of Univ. Lille 1 (M. Pupin, H. Touzet), Univ. Pierre et Marie Curie (L. Noé), ENS Bio (H. Touzet), Université Aix-Marseille (H. Touzet), Univ. Bordeaux 1 (M. Giraud)

[3] CNRS: National Center for Scientific Research

## 8.3. Popularization

- We continued the activity developed on bioinformatics puzzles by our two-months exhibition in 2010 at Palais de la découverte (science museum in Paris). These “puzzles du génome” explain the basics of sequence assembly, RNA secondary structures and phylogenetic reconstruction http://www.lifl.fr/~giraud/puzzles. In 2012, we demonstrated these puzzles to more than 150 high schools pupils (J.F. Berthelot, S. Blanquart, M. Giraud, M. Salson, A. Thomas, M. Pupin).

# 9. Bibliography

## Publications of the year

### Articles in International Peer-Reviewed Journals

[1] A. ABDO, S. CABOCHE, V. LECLÈRE, P. JACQUES, M. PUPIN. *A new fingerprint to predict nonribosomal peptides activity.*, in "Journal of Computer-Aided Molecular Design", October 2012, vol. 26, n$^{o}$ 10, p. 1187-94 [*DOI :* 10.1007/S10822-012-9608-4], http://hal.inria.fr/hal-00750002.

[2] J. ALLALI, C. SAULE, C. CHAUVE, Y. D'AUBENTON-CARAFA, A. DENISE, C. DREVET, P. FERRARO, D. GAUTHERET, C. HERRBACH, F. LECLERC, A. DE MONTE, A. OUANGRAOUA, M.-F. SAGOT, M. TERMIER, C. THERMES, H. TOUZET. *BRASERO: A resource for benchmarking RNA secondary structure comparison algorithms*, in "Advances in Bioinformatics", 2012, accepted for publication (Feb, 2012), http://hal.inria.fr/hal-00647725.

[3] M. BÉCHET, T. CARADEC, W. HUSSEIN, A. ABDERRAHMANI, M. CHOLLET, V. LECLÈRE, T. DUBOIS, D. LERECLUS, M. PUPIN, P. JACQUES. *Structure, biosynthesis, and properties of kurstakins, nonribosomal lipopeptides from Bacillus spp.*, in "Applied Microbiology and Biotechnology", August 2012, vol. 95, n$^{o}$ 3, p. 593-600 [*DOI :* 10.1007/S00253-012-4181-2], http://hal.inria.fr/hal-00749819.

[4] E. KOPYLOVA, L. NOÉ, H. TOUZET. *SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data.*, in "Bioinformatics", October 2012, p. 1-10 [*DOI :* 10.1093/BIOINFORMATICS/BTS611], http://hal.inria.fr/hal-00748990.

[5] M. LÉONARD, L. MOUCHARD, M. SALSON. *On the number of elements to reorder when updating a suffix array*, in "Journal of Discrete Algorithms", February 2012, vol. 11, p. 87-99 [*DOI :* 10.1016/J.JDA.2011.01.002], http://hal.inria.fr/inria-00636066.

[6] E. RIVALS, N. PHILIPPE, M. SALSON, M. LÉONARD, T. COMMES, T. LECROQ. *A Scalable Indexing Solution to Mine Huge Genomic Sequence Collections*, in "ERCIM News", April 2012, vol. 2012, n$^{o}$ 89, p. 20-21, http://hal.inria.fr/lirmm-00712653.

[7] A. SAFFARIAN, M. GIRAUD, A. DE MONTE, H. TOUZET. *RNA Locally Optimal Secondary Structures*, in "Journal of Computational Biology", 2012, vol. 19, n$^{o}$ 10, p. 1120-1133 [*DOI :* 10.1089/CMB.2010.0178], http://hal.inria.fr/hal-00756249.

[8] M. STARTEK, S. LASOTA, M. SYKULSKI, A. BULAK, A. GAMBIN, L. NOÉ, G. KUCHEROV. *Efficient alternatives to PSI-BLAST*, in "bulletin of the polish academy of sciences: technical sciences", 2012, http://hal.inria.fr/hal-00749016.

## International Conferences with Proceedings

[9] M. GIRAUD, R. GROULT, F. LEVÉ. *Detecting Episodes with Harmonic Sequences for Fugue Analysis*, in "ISMIR - International Society for Music Information Retrieval Conference - 2012", Porto, Portugal, 2012, http://hal.inria.fr/hal-00712565.

[10] M. GIRAUD, R. GROULT, F. LEVÉ. *Subject and counter-subject detection for analysis of the Well-Tempered Clavier fugues*, in "International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)", United Kingdom, 2012, p. 661-673, http://hal.inria.fr/hal-00712554.

[11] A. OUANGRAOUA, M. RAFFINOT. *Faster and Simpler Minimal Conflicting Set Identification.*, in "Combinatorial Pattern Matching", Helsinki, Finland, Springer, 2012, vol. 7354, p. 41-55, http://hal.inria.fr/hal-00750028.

[12] A. OUANGRAOUA, K. SWENSON, A. BERGERON. *On the comparison of sets of alternative transcripts*, in "International Symposium on Bioinformatics Research and Applications", Dallas, Germany, Springer, 2012, vol. 7292, p. 201-212, http://hal.inria.fr/hal-00750031.

[13] A. THOMAS, A. OUANGRAOUA, J.-S. VARRÉ. *Genome Halving by Block Interchange*, in "BIOINFORMATICS", Lisboa, Portugal, J. SCHIER, C. M. B. A. CORREIA, A. L. N. FRED, H. GAMBOA (editors), SciTePress, 2012, p. 58-65, http://hal.inria.fr/hal-00749026.

[14] A. THOMAS, A. OUANGRAOUA, J.-S. VARRÉ. *Tandem Halving Problems by DCJ*, in "Workshop on Algorithms in Bioinformatics", Ljubljana, Slovenia, Springer Berlin Heidelberg, 2012, vol. 7534, p. 417-429 [*DOI :* 10.1007/978-3-642-33122-0_33], http://hal.inria.fr/hal-00749019.

## National Conferences with Proceeding

[15] E. KOPYLOVA, L. NOÉ, H. TOUZET. *SortMeRNA: a new software to filter total RNA for metatranscriptomic or RNA analysis*, in "JOBIM - Journées Ouvertes en Biologie, Informatique et Mathématiques - 2012", Rennes, France, 2012, http://hal.inria.fr/hal-00763792.

[16] M. PUPIN, M. SMAIL-TABBONE, P. JACQUES, M.-D. DEVIGNES, V. LECLÈRE. *NRPS toolbox for the discovery of new nonribosomal peptides and synthetases*, in "Journées Ouvertes en Biologie, l'Informatique et les Mathématiques - JOBIM 2012", Rennes, France, F. COSTE, D. TAGU (editors), 2012, p. 89-93, http://hal.inria.fr/hal-00734312.

# References in notes

[17] 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.

[18] 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.

[19] P. GARDNER, R. GIEGERICH. *A comprehensive comparison of comparative RNA structure prediction approaches*, in "BMC Bioinformatics", 2004, vol. 5(140).

[20] 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, p. 180–191 [*DOI :* 10.1007/978-3-540-76336-9_18], http://www.springerlink.com/content/y824l20554002756/.

[21] 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, p. 1–34.

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

[23] 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, p. 384–396 [*DOI :* 10.1007/978-3-642-12683-3_25], http://www.springerlink.com/content/41535x341gu34131/.

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

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

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

[27] P. PETERLONGO, L. NOÉ, D. LAVENIER, G. LES 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, p. 1240-1248 [*DOI :* 10.1007/978-3-540-68111-3], http://www.lifl.fr/~giraud/publis/peterlongo-pbc-07.pdf.

[28] 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.

[29] 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, p. 483–494, http://www.lifl.fr/~noe/files/pp_TCBB09_preprint.pdf.

[30] 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, p. 4350–4359.

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

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