IN PARTNERSHIP WITH:
**CNRS**

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

# Activity Report 2011

# 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, High Performance Computing, RNA Annotation, Nonribosomal Peptides, Genome Rearrangement

BONSAI *is a joint team with LIFL (CNRS UMR 8022, Université Lille 1).*

# 1. Members

**Research Scientists**
Hélène Touzet [Team leader, Senior Researcher CNRS, Maternity leave from February to July, HdR]
Samuel Blanquart [Junior Researcher INRIA]
Mathieu Giraud [Junior Researcher CNRS]
Aïda Ouangraoua [Junior Researcher INRIA]

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

**External Collaborators**
Louise Ott [Ingénieur, INRA, until October 2011]
Laurie Tonon [Expert Engineer, INRIA, until July 2011]

**Technical Staff**
Jean-Frédéric Berthelot [IJD, INRIA]

**PhD Students**
Azadeh Saffarian [MESR fellowship, from November 2007]
Antoine Thomas [MESR fellowship, from October 2010]
Tuan Tu Tran [INRIA CORDI fellowship, from September 2009]
Evguenia Kopylova [ANR grant, from December 2010]

**Post-Doctoral Fellow**
Ségolène Caboche [ATER, Université Lille 1, until August 2011]

**Visiting Scientist**
Patrick Meyer [Université Libre de Bruxelles, October 2011, two weeks]

**Administrative Assistant**
Sandrine Catillon [INRIA]

# 2. Overall Objectives

## 2.1. Presentation

The team BONSAI has been re-created on January 1, 2011, and is an evolution of the INRIA-LIFL team Sequoia, which was created in 2007. The scientific focus of Bonsai is still very much the same as the one of Sequoia. We work in computational biology, and more specifically o n algorithms for biological sequences analysis. Several topics of Bonsai were already present in Sequoia: Noncoding RNA analysis and non ribosomal peptide synthesis. We also work on further lines of research: Algorithms for Next Generation Sequencing and comparison of sequences at genome scale taking into account rearrangements. These lines of research find their source in the development of new sequencing technologies and the increasing availability of complete genome sequence data. They are supported by strategical collaborations, and they also reinforce the expertise of the team in sequence analysis and genome annotation. The main goal of BONSAI is to define appropriate combinatorial models and efficient algorithms for large-scale sequence analysis in molecular biology.

From a historical perspective, research in bioinformatics started with string algorithms designed for the comparison of sequences. Bioinformatics became then more diversified, accompanying the emergence of new high-throughput technologies. By analogy to the living cell itself, bioinformatics is now composed of a variety of dynamically interacting components forming a large network of knowledge: systems biology, proteomics, text mining, phylogeny, structural genomics, etc. **Sequence analysis** 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. From a computational point of view, NGS gives rise to many new challenging problems in computational biology.

The completion of sequencing projects in the past few years also teaches us that the functioning of the genome is more complex than expected, which makes genomic sequence annotation a difficult task. Originally, genome annotation was mostly driven by protein-coding gene prediction. It is now widely recognized that deciphering gene regulation is an essential task to understand the function of a protein. **Noncoding RNA** genes also play a key role in many cellular processes. Besides proteins encoded in the genome, there exist other, more specific, processes such as **nonribosomal peptide synthesis** that produces small peptides not going through the central dogma. At a higher level, **genome organization and rearrangements** are also a source of complexity and have a high impact on the course of evolution.

All above-mentioned biological phenomena together with big volumes of new sequence data and new hardware provide a number of new challenges to bioinformatics, both on modeling the underlying biological mechanisms and on efficiently treating the data. BONSAI is a fully interdisciplinary project whose scientific core is the design of efficient algorithms for the analysis of biological macromolecules. We focus on four main themes: high-throughput sequence analysis, noncoding RNAs, nonribosomal peptides, and genome rearrangements. See Figure 1 for an illustration. Most of 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 models and algorithms

Our research is driven by biological questions. At the same time, we have in mind to develop well-founded models and algorithms. This is essential to guarantee the universality of our results. Our main background comes from *combinatorial discrete models and algorithms*. Biological macromolecules are naturally modelled by various types of discrete structures: String, trees, and graphs, etc.

String algorithms is an established research subject of the team. We have been working on spaced seed techniques for several years [22], [23], [24], [33], [35], [27], [26].The whole technique is implemented and made available in the YASS software for DNA sequence alignment together with the tools implemented to design seeds [28] (see Section 4.).

Members of the team have also a strong expertise in text indexing data structures that are widely-used for the analysis of biological sequences because they allow a data set to be stored and queried efficiently. We proposed an optimal neighborhood indexing for protein similarity search [34] and compressed index structures for DNA sequences [37], [36].

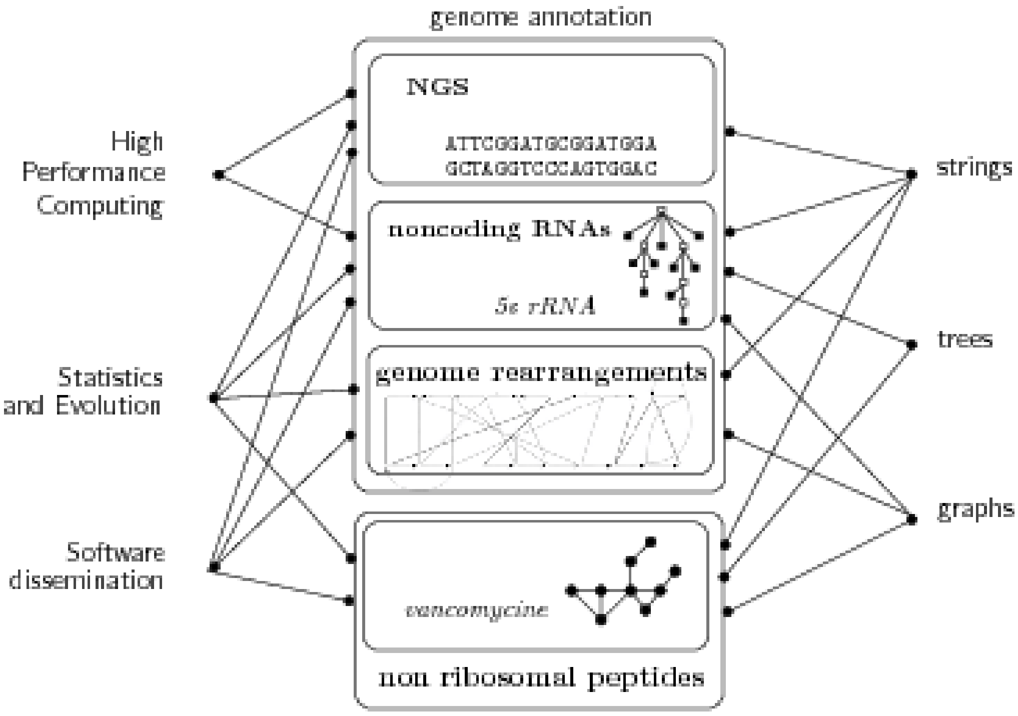

*Figure 1.* BONSAI *at a glance: Research topics and common features*

Ordered trees and graphs naturally arise when dealing with structural RNAs. Our knowledge in this field allowed us to make several significant contributions to RNA bioinformatics on the past few years. First, we proposed a new method for RNA structure inference, implemented in a program called CARNAC, Second, we worked on theoretical models for RNA comparison, which led to substantial advances on tree edit distance algorithms [20], [38], [31], tree models [30], [29] and comparison of arc-annotated sequences [18], [17].

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.

Nonribosomal peptides representation also uses graphs: Nonribosomal peptides are small molecules that have a branching or cyclic structure. We developed several efficient algorithms to compare NRP molecules represented as non-oriented labeled graphs [19].

## 3.2. High-performance computing

*High-performance computing* is another tool that we will use to achieve our goals. It covers several paradigms: grids, single-instruction, multiple-data (SIMD) instructions, graphics cards (GPU). In a near future, processors may offer tens or hundreds of cores with large vector units, combining again several levels of parallelism. Libraries like CUDA and OpenCL also facilitate the use of these manycore processors. This new hardware architecture brings 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*. Probabilistic models indeed naturally appear in many of our research projects. 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. Moreover, probabilistic models described according to the Bayesian framework allow to bypass, by using MCMC sampling methods, some limitations resulting from complex mathematical integrations over parameter space. Bayesian models and their MCMC sampling allow to approximate probability distributions over parameters and to describe more biologically relevant models. These methods are applied to the genome rearrangement application domain.

# 4. Application Domains

## 4.1. Application Domains

- Sequence processing for Next Generation Sequencing

  In the last years, sequencing techniques experienced remarkable advances with *next generation sequencing* (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, ...

- Noncoding RNAs

  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. They also need complex evolution models.

- Genome rearrangements

  Genome organization is also a source of complexity in genome. 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, computing ancestral genomes and paths transforming these genomes into several descendant genomes.

- Nonribosomal peptides

  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 [25]. 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/

Objective : YASS is an open source 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.

YASS is frequently used, it always receives more than 300 web queries per month (excluding INRIA and Univ-Lille1 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/

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 [21].

## 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/

The TFM suite is a set of tools for analysis of transcription factor binding sites. locating and analyzing transcription factor binding sites using 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 developped in 2011.*
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 considering locally optimal structures. Locally optimal structures are thermodynamically stable 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. The tool generates all locally optimal structures in a given sequence. Moreover, REGLISS can be used to explore the neighborhood of structures through an energy landscape graph.

## 5.5. RNAspace – a platform for noncoding RNA annotation

*Actively developped in 2011.*
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/

RNAspace is an open source platform born from a national collaborative initiative. Its goal is to develop and integrate functionalities allowing structural and functional noncoding RNA annotation (see Section 6.2): http://www.rnaspace.org, and it is distributed under the GPL licence. The project has been awarded by the national IBISA label in autumn 2009[1].

## 5.6. CGseq – a toolbox for comparative analysis

*Actively maintained in 2011.*
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/

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

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

## 5.7. Biomanycores.org – a community for bioinformatics on manycore processors

*Actively developped in 2011.*
Software self-assessment: **A-3up4**, **SO-2**, **SM-2**, **EM-3**, **SDL-4up5**, DA-4, CD-4, MS-4, TPM-4

Software web site : http://www.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 is supported by a national ADT[2] between BONSAI, SYMBIOSE (CRI Rennes) and DOLPHIN (CRI Lille). This ADT started in October 2010 and led to the hiring of J.-F. Berthelot (IJD).

In the first year of the ADT, J.-F. Berthelot redesigned and rewrote almost all the existing code. The code base is now stable. He worked on the documentation and on various software engeneering aspects such as continuous integration. The second year of the ADT will focus on integrating more applications and targeting bioinformaticians users.

## 5.8. Norine – a resource for nonribsomal peptides

*Actively developped in 2011.*
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/

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.

Project management: 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.

Users community: Since its creation in 2006, Norine has gained a universal recognition as the unique database dedicated to non-ribosomal peptides because of its high quality and manually curated annotations. 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.

Improvements: This year, the source code has been reorganised by Laurie Tonon, a SED engineer, to use model view controller software architecture, implemented with Struts2.

## 5.9. 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.

[2] ADT (Action for Technological Development) is an INRIA internal call

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). We plan to improve our library in the forthcoming months with the help of Master's students.

# 6. New Results

## 6.1. High-throughtput sequence processing

- We published a book chapter on bioinformatics algorithms for GPU/manycore processors [16].
- Within the PhD of T. T. Tran, we proposed a new bit-parallel algorithm, extension of [39], as well as a new indexing structure adpated to GPUs [13].
- We proposed a new index structure specifically designed for reads produced by high throughput sequencers. It can deal both with variable or fixed length reads and can index reads in usually less memory than other classical solutions. This index has been implemented and is available online [10].
- We characterised the number of elements to be reordered when updating a full-text index such as a suffix array. We finally concluded that this number tends to be poly-logarithmic in the input length for DNA sequences [5].

## 6.2. Noncoding RNAs

- A. Saffarian defended her thesis on November 16. Within her thesis, we obtained two results :
  - 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. A journal article is in revision to *Journal of Computational Biology*.
  - We also proposed an algorithm to match a multi-structure of RNA against a sequence. A multi-structure gather several RNA structures, as real or putative structures on a same sequence, or as similar structures in a family of RNAs. A journal article was submited to *Algorithms for Molecular Biology*.
- We published an algorithm for the comparison RNA secondary structures represented as nested arc-annotated sequences [7].
- The non-coding RNA collaborative annotation platform RNAspace is made available to the community, and published in *RNA* [3].

## 6.3. Genome rearrangements

- A. Thomas has started his PhD on September 2011. We already obtained two results:
  - 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 has been accepted in Bioinformatics 2012.
  - We also introduced and study a new combinatorial problem, a biological phenomenon that apparently associates a significant proportion of segmental duplications in mammalians, drosophilas and bacterias to breakpoints in rearrangement events. called the Genome Dedoubling Problem. It consists in finding a minimum length rearrangement scenario required to transform a genome with duplicated segments into a non-duplicated genome such that duplications are caused by rearrangement breakpoints. We introduced new graph data structures to solve these problems. This work was presented at RECOMB'CG 2011 [11]

- We designed and applied new algorithms for infering ultra-perfect evolution scenarios for Drosophila and mammals species [6].
- We implemented and applied the algorithm of [32] for the reconstruction of species tree from gene trees of Fungal and eukaryotes species [8].
- We proposed a new reconstruction of the architecture of the ancestral amniote genome based on the detection and assembly of ancestral genomic features conserved in extant species [9].

## 6.4. Non-ribosomal peptides

- A new database, called Doris (for Database Of non-RIbosomal Synthetases), has been created to extend the information we provide about non-ribosomal peptides to their producing enzymes, the non-ribosomal peptide synthetases. For the moment, a first version of the web interface has been developed by Louise Ott, an engineer from Lille1 University. More than 400 enzymes have been automatically extracted from general databases.
- A collaboration started with members of Orpailleur EPI to design a semi-automatic process to collect non-ribosomal synthetases (NRPSs). We already start adapting MODIM, a generic tool developed by Orpallieur EPI to collect and integrate data extracted from various web sources, to the specific needs of NRPSs.
- As mentioned in the software section, the source code has been reorganised by Laurie Tonon, a SED engineer, to use model view controller software architecture, implemented with Struts2.

# 7. Partnerships and Cooperations

## 7.1. Regional Initiatives

Bioinformatics is a multidisciplinary discipline by nature and our work relies on collaborations with several biological research groups.

- The project on *nonribosomal peptide synthesis* is based on a collaboration with the ProBioGEM laboratory (*Laboratoire des Procédés Biologiques Génie Enzymatique et Microbien*), headed by Pr. Dhulster, University Lille 1. This laboratory develops methods to produce and extract active peptides in agriculture or food. The PhD work of Ségolène Caboche defended in 2009 was co-supervised by Valérie Leclère from ProBioGem. A PhD work started on this subject in 2008: Aurélien Vanvlassenbroeck is working at ProBioGEM and is co-supervised by Maude Pupin.
- We collaborate with the *Laboratoire de Génétique et Évolution des Populations Végétales* (UMR CNRS 8016), Université de Lille 1 on the study of genomic rearrangements in the mitochondrial genome of higher plants. The goal is to identify evolutionary forces and molecular mechanisms that modeled the present diversity of mitochondrial genome at the species level, and in particular potentially active recombination sequences that have been used in the course of time. Data is acquired thanks to Genoscope projects (in beet and silene). A PhD work defended in 2010 by Aude Darracq was co-supervised by Pascal Touzet from GEPV.
- 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.
- Our team is a member of the *PPF Bioinformatique*. This is an initiative of the University Lille 1 that coordinates public bioinformatics activities at the local level for the period 2010-13.
- We collaborate with the INSERM unit 800, Université Lille 1, to infer scenario for the creation of new exons and new alternative splicings during the evolution TRPM8 enzyme in Human.

## 7.2. National Initiatives

- ANR MAPPI (2010-2013). ANR Mappi (2010-2013): National funding from the French Agency Research (call *Conception and Simulation*). This project involves four partners: LIAFA (Université Paris 7), SYMBIOSE (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*.
- RNAspace. This project conducted in collaboration with INRA Toulouse benefited from a grant of RNG-Renabi, national network for bioinformatics, and is one of the topics of the project RENABI-IFB (national call *infrastructures Biologie Santé*).
  Project web site : http://www.rnaspace.org/
- A collaboration with the PlasmoExplore ANR project (Laboratoire d'Informatique et Microélectronique de Montpellier) led to a publication [2]. We propose in this paper a new evolutionary scenario for the evolution of Plasmoduim falciparum, the major agent of malignant malaria. Results issued from the collaboration with the ERC founding PopPhyl are in preparation (Institut des Sciences de l'Evolution de Montpellier).
- The following scientists were invited in 2011 to give a talk at the team seminar:
  Patrick Meyer (Université Libre de Bruxelles), José Gualberto (Institut de Biologie Moléculaire des Plantes, Strasbourg), Robert Giegerich (Université Bielefeld).

## 7.3. European Initiatives

- PHC Procope PARALLEL-ADP (2010-2011), bilateral cooperation project with U. Bielefeld (R. Giegerich, P. Steffen, Germany). The goal is to work on a generic parallelization on the ADP (algebraic dynamic programming) methodology. In this context, R. Giegerich spent several days in Lille in November.

## 7.4. International Initiatives

### *7.4.1. INRIA International Partners*

A collaboration with the Université du Québec à Montréal (UQAM) and Simon Fraser University (SFU, Vancouver) on gene evolution, genomic rearrangement and ancestral genome reconstruction led to three publications this year [8], [6], [9].

### *7.4.2. Visits of International Scientists*

Patrick Meyer (Université Libre de Bruxelles, October 2011, two weeks).

Robert Giegerich (University of Bielefeld, Germany, November 2011, three days).

# 8. Dissemination

## 8.1. Animation of the scientific community

### *8.1.1. Administrative activities*

- 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 take part yearly to the organization of a annual national workshop on sequence analysis and bioinformatics. This year, the workshop is hosted in Lille.

- H. Touzet is a 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 biology sciences.
- Member of the INRIA evalution commitee (M. Giraud)
- Scientific secretary of the Gilles Kahn PhD award commitee (M. Giraud)
- Member of ITMO Genetics, Genomics and Bioinformatics of AVIESAN (H. Touzet)
- The team is in charge of the PPF *Bioinformatique*. This is an initiative of Université Lille 1 that coordinates 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. In this context, we organized three one-day workshops, that gathered around 60 people each: *high performance computing for biology*, *text mining in biology*, *Analysis of NGS data*.
- Head of PPF bioinformatics – University Lille 1 (H. Touzet)
- Head of CIB, 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)
- Coordinator for the RAweb 2011 of INRIA Lille Nord Europe (A. Ouangraoua)
- Member of Cordis-Postdoc committee for LNE center (M Pupin)
- Member of UFR IEEA council (M. Pupin)
- Member of Polytech'Lille council (S. Janot)
- Member of the LIFL Laboratory council (H. Touzet)
- Member of hiring committee *(jury d'audition)* of IUT de Strasbourg (L. Noé), Université Lille 1 (M Pupin), Université Bordeaux 1 (M Giraud), INRIA (M Giraud), Université Montpellier 2 (M. Salson)

### 8.1.2. Project reviewing activities

- Reviewer for the Research Foundation Flanders (FWO) (J.-S. Varré)
- Reviewer for CNRS PEPS and PEPII programs (H. Touzet)

### 8.1.3. PhD theses, and HdR committees

- Member of the thesis committee of Anne-Laure Gaillard, Univ. Bordeaux 1 (M. Giraud),
- Member of the thesis committee of Kunthea Phok, INRA Toulouse, Anouar Ben Hassena, Université Rennes 1, and habilitation committee of Eric Tannier, Université Lyon 1 (H. Touzet)

### 8.1.4. Editorial and article reviewing activities

- Program committee of WABI 2011 (H. Touzet), Recomb-CG 2011 (A. Ouangraoua), PPAM-PBC 2011 (M. Giraud) WEPA 2012 (M. Giraud)
- Reviewer for the journals Bioinformatics (L. Noé, J.-S. Varré), BMC Bioinformatics (L. Noé, H. Touzet), BMC Research Notes (M. Giraud), Theoretical Computer Science (L. Noé, M. Salson), Transactions on Computational Biology and Bioinformatics (M. Giraud, J.-S. Varré), Nucleic Acids Research (J.-S. Varré), Computers in biology and Medicine (H. Touzet), BMC Evolutionary Biology and Molecular Biology and Evolution (S. Blanquart)
- Reviewer for the conferences ACM-SAC BIO2011 (L. Noé, J.-S. Varré, H. Touzet), CPM 2011 (L. Noé), EuroPar 2011 (M. Giraud), SPIRE 2011 (M. Giraud, L. Noé), STACS 2011 (M. Giraud), PSC 2011 (M. Salson), IC3 2011 (J.-S. Varré),

[3] CNRS: National Center for Scientific Research

## 8.2. 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).

### *8.2.1. Licence (bachelor)*

Teaching in computer science (University of Lille 1, unless otherwise stated)

- *Programming (OCaml)*, 50h, L2 (second year of bachelor) (J.-S. Varré)
- *Algorithms (Ada)*, 58h, L3 (third year of bachelor) (L. Noé)
- *Networks*, 36h, L3 (third year of bachelor) (L. Noé)
- *System*, 36h, L3 (third year of bachelor) (L. Noé)
- *Algorithms and Data Structures*, 60h, L3 (third year of bachelor) (J.-S. Varré)
- *Introduction to Programming (OCaml)*, 54h, L1 (first year of bachelor) (M. Salson)
- *Coding and information theory*, 36h, L2 (second year of bachelor) (M. Salson)
- *C programming*, 42h, L3 (third year of bachelor) (M. Salson)
- *Imperative Programming*, 48h, L1 (first year of bachelor) (M. Giraud)
- *Computation and Architecture*, 36h, L1 (first year of bachelor) (M. Giraud, Univ. cath. of Lille)
- *Introduction to programming*, 50h, first year of engineering school (L3) (S. Janot)
- *Introduction to Databases*, 30h, first year of engineering school (L3) (S. Janot)
- *Programming (OCaml)*, 36h, L1 (first year of bachelor) (M. Pupin)
- *Professional project*, 18h, L3 (third year of bachelor) (M. Pupin)
- *Algorithmics*, second year of bachelor, 30h, L2 (second year of bachelor) (A. Saffarian)

### *8.2.2. Master*

Lectures on bioinformatics, University of Lille 1.

- *Bioinformatics*, 54h, M1 (first year of master in genomics and proteomics) (L. Noé)
- *Bioinformatics*, 24h, M1 (first year of master in genomics and microbiology) (A. Ouangraoua)
- *Bioinformatics*, 36h, M2 (second year of master in cellular and molecular engineering) (J.-S. Varré)
- *Applications and Algorithms in bioinformatics*, 21h, M1 (first year of master in computer science) (S. Blanquart)
- *Introduction to Programming (JAVA)*, 30h, M1 (first year of master) (M. Pupin)
- *Individual project*, organiser, M1 (fisrt year of master in computer science) (L. Noé)
- *Databases*, 12h, second year of engineering school (M1) (S. Janot)
- *Introduction to Artificial Intelligence*, 25h, second year of engineering school (M1) (S. Janot)
- *Artificial Intelligence and Constraint programming*, 25h, third year of engineering school (M2) (S. Janot)
- *Final project*, organiser, third year of engineering school (M2) (S. Janot)

### 8.2.3. PhD

- *Azadeh Saffarian*, Prediction and pattern matching algorithms for RNA multi-structures, Université Lille 1, November 16, 2011, co-directed by H. Touzet and M. Giraud.
- *Tuan Tu Tran*, Massively parallel algorithms and data structures for bioinformatics, Université Lille 1, in progress, co-directed by J-S. Varré and M. Giraud.
- *Antoine Thomas*, Algorithms for genome rearrangement with duplications, Université Lille 1, in progress, co-directed by J-S. Varré and A. Ouangraoua.
- *Evguenia Kopylova*, New algorithmic and bioinformatic approaches for the analysis of data from next-generation sequencing, Université Lille 1, in progress, co-directed by H. Touzet and L. Noé.

## 8.3. Popular science

- 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 2011, we demonstrated these puzzles to more than 350 pupils in 7 high schools of the region (J.F. Berthelot, S. Blanquart, M. Giraud, L .Ott, A. Saffarian, M. Salson).

# 9. Bibliography

## Références à compléter (Corrigez dans Hal et regénérez votre trame dans 24h)

### A modifier : x-proceedings={yes|no},

[1] A. GAMBIN, S. LASOTA, M. STARTEK, M. SYKULSKI, L. NOÉ, G. KUCHEROV. *Subset seed extension to Protein BLAST*, in "Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2011)", SciTePress, 2011, p. 149-158 [*DOI :* 10.5220/0003147601490158], http://hal.inria.fr/inria-00609791/en.

## Publications of the year

### Articles in International Peer-Reviewed Journal

[2] S. BLANQUART, O. GASCUEL. *Mitochondrial genes support a common origin of Plasmodium falciparum and relatives with rodent malaria parasites.*, in "BMC Evolutionary Biology", March 2011, http://hal.inria.fr/inria-00636084/en.

[3] M.-J. CROS, A. DE MONTE, J. MARIETTE, P. BARDOU, B. GRENIER-BOLEY, D. GAUTHERET, H. TOUZET, C. GASPIN. *RNAspace.org: An integrated environment for the prediction, annotation, and analysis of ncRNA*, in "RNA", September 2011, vol. 17, p. 1947-1956, http://rnajournal.cshlp.org/content/17/11/1947.long.

[4] M. GIRAUD, J.-S. VARRÉ. *Parallel Position Weight Matrices Algorithms*, in "Parallel Computing", 2011, vol. 37, p. 466-478 [*DOI :* 10.1016/J.PARCO.2010.10.001], http://hal.inria.fr/hal-00623404/en.

[5] M. Léonard, L. Mouchard, M. Salson. *On the number of elements to reorder when updating a suffix array*, in "Journal of Discrete Algorithms", January 2011 [*DOI :* 10.1016/j.jda.2011.01.002], http://hal.inria.fr/inria-00636066/en.

[6] A. Ouangraoua, A. Bergeron, K. Swenson. *Theory and practice of ultra-perfection*, in "Journal of Computational Biology", 2011, vol. 18, n$^{o}$ 9, p. 1219-1230, http://hal.inria.fr/inria-00635033/en.

[7] A. Ouangraoua, V. Guignon, S. Hamel, C. Chauve. *A new algorithm for aligning nested arc-annotated sequences under arbitrary weight schemes*, in "Theoretical Computer Science", 2011, vol. 412, n$^{o}$ 8-10, p. 753-764, http://hal.inria.fr/inria-00635043/en.

[8] A. Ouangraoua, K. Swenson, C. Chauve. *A 2-Approximation for the Minimum Duplication Speciation Problem*, in "Journal of Computational Biology", 2011, vol. 18, n$^{o}$ 9, p. 1041-1053, http://hal.inria.fr/inria-00635025/en.

[9] A. Ouangraoua, E. Tannier, C. Chauve. *Reconstructing the architecture of the ancestral amniote genome*, in "Bioinformatics", 2011, n$^{o}$ 19, p. 2664-2671, http://hal.inria.fr/inria-00635016/en.

[10] N. Philippe, M. Salson, T. Lecroq, M. Léonard, T. Commes, E. Rivals. *Querying Large Read Collections in Main Memory: A Versatile data Structure*, in "BMC Bioinformatics", Jun 2011, vol. 12, 242+ [*DOI :* 10.1186/1471-2105-12-242], http://hal.inria.fr/lirmm-00632958/en.

[11] A. Thomas, J.-S. Varré, A. Ouangraoua. *Genome Dedoubling by DCJ and Reversal*, in "BMC Bioinformatics", 2011, vol. 12, n$^{o}$ Suppl 9, S20, Proceedings of the Ninth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics, http://hal.inria.fr/inria-00635003/en.

### International Conferences with Proceedings

[12] F. Levé, R. Groult, G. Arnaud, C. Séguin, R. Gaymay, M. Giraud. *Rhythm extraction from polyphonic symbolic music*, in "12th International Society for Music Information Retrieval Conference (ISMIR 2011)", United States, 2011, p. 375-380, http://hal.inria.fr/hal-00636058/en.

[13] T. T. Tran, M. Giraud, J.-S. Varré. *Bit-Parallel Multiple Pattern Matching*, in "Parallel Processing and Applied Mathematics / Parallel Biocomputing Conference (PPAM / PBC 11)", Torun, Poland, 2011, http://hal.inria.fr/inria-00637227/en.

### Conferences without Proceedings

[14] J.-F. Berthelot, C. Deltel, M. Giraud, S. Janot, L. Jourdan, D. Lavenier, H. Touzet, J.-S. Varré. *biomanycores.org: a repository of interoperable open-source code for many-core bioinformatics*, in "JOBIM 2011", Paris, France, July 2011, http://hal.inria.fr/hal-00637847/en.

[15] M. Giraud, S. Janot, J.-F. Berthelot, C. Deltel, L. Jourdan, D. Lavenier, H. Touzet, J.-S. Varré. *Biomanycores, open-source parallel code for many-core bioinformatics*, in "Bioinformatics Open Source Conference (BOSC 2011)", Vienne, Austria, 2011, http://hal.inria.fr/inria-00623390/en.

### Scientific Books (or Scientific Book chapters)

[16] J.-S. VARRÉ, B. SCHMIDT, S. JANOT, M. GIRAUD. *Manycore high-performance computing in bioinformatics*, in "Advances in Genomic Sequence Analysis and Pattern Discovery", L. ELNITSKI, H. PIONTKIVSKA, L. R. WELCH (editors), World Scientific, 2011, chapter 8, http://hal.inria.fr/hal-00563408/en.

## 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] G. BLIN, H. TOUZET. *How to Compare Arc-Annotated Sequences: The Alignment Hierarchy*, in "13th International Symposium on String Processing and Information Retrieval (SPIRE)", Lecture Notes in Computer Science, Springer Verlag, 2006, vol. 4209, p. 291–303 [*DOI :* 10.1007/11880561_24], http://www.springerlink.com/content/4k37q116j2720832/.

[19] 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.

[20] S. DULUCQ, H. TOUZET. *Decomposition algorithms for the tree edit distance problem*, in "Journal of Discrete Algorithms", 2005, p. 448-471, http://dx.doi.org/10.1016/j.jda.2004.08.018.

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

[22] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *Multi-seed lossless filtration*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", January-March 2005, vol. 2, n$^{o}$ 1, p. 51–61.

[23] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *A unifying framework for seed sensitivity and its application to subset seeds*, in "Journal of Bioinformatics and Computational Biology", 2006, vol. 4, n$^{o}$ 2, p. 553–569 [*DOI :* DOI:10.1142/S0219720006001977], http://www.worldscinet.com/jbcb/04/0402/S0219720006001977.html.

[24] 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/.

[25] 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.

[26] 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/.

[27] 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/.

[28] L. NOÉ, G. KUCHEROV. *YASS: enhancing the sensitivity of DNA similarity search*, in "Nucleic Acid Research", 2005, vol. 33, p. W540-W543.

[29] 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.

[30] 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.

[31] 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.

[32] A. OUANGRAOUA, K. SWENSON, C. CHAUVE. *An approximation algorithm for computing a parsimonious first speciation in the gene duplication model*, in "Proceedings of RECOMB-Comparative Genomics, LNBI 6398", 2010, vol. 6398, p. 290-302.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] S. WU, U. MANBER. *Fast Text Searching Allowing Errors*, in "Communications of the ACM", october 1992, vol. 35, n$^{o}$ 10, p. 83–91.