INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team sequoia2

Algorithms for large scale sequence analysis

Lille - Nord Europe

Theme : Computational Biology and Bioinformatics

Activity Report

2010

Table of contents

SEQUOIA2 *is a joint team with LIFL (CNRS UMR 8022, Université Lille 1).*

1. Team

Research Scientists

Hélène Touzet [Team leader, Senior Researcher CNRS, HdR]
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, on leave at INRIA until September 2010]
Mickaël Salson [Associate Professor, Université Lille 1, from September 2010]
Ségolène Caboche [ATER, Université Lille 1, from September 2010]

Technical Staff

Jean-Frédéric Berthelot [IJD, INRIA, from November 2010]
Antoine de Monte [Ingénieur, INRA, until September 2010]
Laurie Tonon [IJD, INRIA, until September 2010]

PhD Students

Aude Darracq [MESR fellowship, from October 2007]
Marta Girdea [INRIA CORDI fellowship, from October 2007]
Azadeh Saffarian [MESR fellowship, from November 2007]
Antoine Thomas [MESR fellowship, from October 2010]
Tuan Tu Tran [INRIA CORDI fellowship, from September 2009]

Administrative Assistant

Sandrine Catillon [INRIA]

2. Overall Objectives

2.1. Presentation

The main goal of SEQUOIA2 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,... Sequence analysis remains a central node in this interconnected network, and it is the heart of the SEQUOIA2 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.

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 *non-ribosomal peptide synthesis* that produces small peptides not going through the central dogma. At a higher level, genome organization and large-scale genome 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. SEQUOIA2 is a fully interdisciplinary project. 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/

2.2. Highlights

- Our last paper on non-ribosomal peptides [2] is cited in the Journal Highlights of the Microbe magazine edited by the American Society for Microbiology (see section 6.5.1)
- We organized a two-month exhibition on bioinformatics puzzles in Palais de la Découverte (science museum in Paris, see section 8.3)

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, ...

String algorithms is an established research subject of the team. We have been working on spaced seed techniques for several years [22], [23], [24], [32], [30]. The whole technique is implemented and made available in the YASS software for DNA sequence alignment together with the tools implemented to design seeds [26] (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 [31] and compressed index structures for DNA sequences [34], [33].

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], [36], [29], tree models [28], [27] and comparison of arc-annotated sequences [18], [17].

String, trees and graphs are also present to study genomic rearrangements: Neighborhoods of genes can be modelled by oriented graphs, genomes as permutations, sequences or trees. *Non-ribosomal peptides*, is also concerned with graphs: Non-ribosomal 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, ...

4. Application Domains

4.1. Application Domains

- Sequence processing for Next Generation Sequencing

  In the last three 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 analyze large sets of genomes, computing ancestral genomes and paths transforming these genomes into several descendant genomes.

- Cis-regulatory motifs

  Another important aspect of the analysis of genomes concerns gene regulation. Gene expression is controlled at several levels: mRNA transcription, mRNA processing, protein synthesis, post-translational modifications, RNA degradation. Genome analysis can help to elucidate the very first step in this chain: transcriptional regulation. Transcription of a gene is controlled by regulatory proteins – such as transcription factors (TFs) – that bind to the DNA. This protein/DNA interaction requires a binding site whose sequence pattern is more or less specific to each TF. Identification of transcription factor binding sites (TFBSs) is a notoriously difficult task because motifs corresponding to TFBSs have a very low information content: they are usually short (around 5-15 bases) and degenerate. Modeling, identification and analysis of TFBSs is one of major bioinformatics challenges.

- Non-ribosomal peptides

  Non-ribosomal 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 non-ribosomal 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 non-ribosomal 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.
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.
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.

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 [10].

5.4. Regliss – RNA locally optimal structures

Actively developped in 2010.
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 2010.
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 6.3): 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 developed in 2010.
CG-seq is a toolbox to identify functional regions in a genomic sequence by comparative analysis using multi-species comparison.

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

Actively developped in 2010.
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.

We will pursue its development in active collaboration with SYMBIOSE (CRI Rennes) and DOLPHIN (CRI Lille) with the support of a national ADT[2] that started in October 2010. This support allows us to have a software engineer for two years, and we just hired J.-F. Berthelot on this position.

5.8. Norine – a resource for nonribsomal peptides

Actively developped in 2010.
Norine is a public computational resource that contains a database of NRPs coupled with dedicated tools. The web interface, including the tools for comparing NRPs, is developed in JSP (JavaServer Pages) and totalizes around 13,000 lines of code. It also includes visualization tools, such as a 2D graph viewer and editor for peptides.

[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

Norine is queried from all around the world. 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 biolgy laboratories work on non-ribosomal peptides (NRPs) or on their synthetases. Interface and functionalities of Norine have been significantly improved this year. See Section 6.5.2.

6. New Results

6.1. High-throughtput sequence processing

6.1.1. Seed-based sequence comparison and mapping

- Within the PhD of M. Gîrdea, we proposed a new method to efficiently design *position restricted seeds* in a *lossless* or a *lossy* framework. This work extends and unifies previous works done on seed design, and focus on NGS technologies. This method was thus applied in the very interesting SOLiD next generation sequencing technology, where several fitted models were proposed and set in the lossless and lossy frameworks [12], [7]. This work has been implemented in a prototype software named SToRM

6.1.2. High-performance computing for bioinformatics

- We have been invited to write a book chapter on bioinformatics algorithms for GPU/manycore processors [16].
- Within the PhD of T. T. Tran, we drafted an indexing structure for GPUs. This is an ongoing work.
- M. Giraud spent one month (April 2010) in U. Bielefeld to continue the work on parallelization of Algebraic Dynamic Programming [35].

6.2. Cis-regulatory elements

- We proposed a new version of our software TFM-Explorer for the analysis of regulatory regions in eukaryotic co-regulated genes. We improved the algorithm enabling detection of cis-regulatory modules, enhanced visualization tools and added new species models [10].
- We finished the work on parallel algorithms for Position Weight Matrices that was begun in 2009. In 2010, we improved the methods for the scan of multiple Position Weight Matrices [5].

6.3. Noncoding RNAs

6.3.1. RNA Suboptimal Structures

- Within the PhD of A. Saffarian, 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 has been submitted.

6.3.2. RNA pattern matching

- We proposed a new algorithm to identify putative occurrences of a RNA in a genomic sequence allowing for alternative foldings. Experimental validation is underway.

6.4. Genome rearrangements

6.4.1. Combinatorial models and algorithms

- On sorting duplicated genomes : we designed new algorithms for sorting duplicated genomes into tandem duplication configuration. This work was achieved during the Master internship of A. Thomas from February to April 2010. The valuation of these results is underway and the work continues through the Ph.D. of Antoine started in October 2010.
- Representation of rearrangement scenarios : we introduced new useful representations of rearrangement scenarios as combinatorial objects (integer sequences, trees). These results led to new algorithms for the uniform generation of random rearrangement scenarios [8].
- Ultra-perfect rearrangement scenarios : We introduced a new model called Ultra-perfection" for the evolution of genomes. Ultra-perfect calls for the conservation of co-located genes in evolution scenarios between not only extant, but also ancestral genomes [13].
- Reconstruction of ancestral genomes : For the reconstruction of ancestral genomes, we designed and used new computational methods allowing to deal with the hardness caused by whole-genome duplication events. We applied these methods to the reconstruction of the evolutionary history of yeasts genomes [3].

6.4.2. Analysis of plant mitochondrial genomes

- On analysis of maize mitochondrial genomes : We proposed a methodology to take into account duplicated genes according tandem duplication as the main duplication event and computed a rearrangement phylogeny on plant mitochondrial genomes for the first time [4].
- On analysis of beet mitchondrial genomes : We sequenced 5 mitochondrial genomes in beet and proposed congruent phylogenies using sequences of genes or rearrangements (PhD Thesis of Aude Darracq [1]). A journal article has been submitted.

6.5. Non-ribosomal peptides

6.5.1. Insights on Non-ribosomal Peptides

- We performed the first large-scale statistical analysis of the non-ribosomal peptides, revealing at least 500 unique monomers incorporated in these compounds. We showed differences between the structural properties of bacterial and fungal NRPs; that NRPs synthesized by bacteria and those isolated from metazoa appear similar, supporting the hypothesis that peptides isolated from sponges are synthesized by their symbiotic bacteria. We observed that some monomeric structures indicate specificity to certain classes of biological activities. This work has been published in the Journal of bacteriology[2] and highlighted by the American Society for Microbiology in their Microbe magazine [3]

6.5.2. New version of Norine

- The interface of Norine has been notably improved to facilitate its querying (possibility to do query refinements from graphical output, enhanced version of the structure editor) and the read of the results (graphical output, peptide and monomers lists are hierarchically represented).
- New peptides have been added and a monomer classification has been created.
- Crosslinks with PDB are integrated (14 PDB entries are linked to Norine and 20 Norine entries are linked to PDB), more are coming.

7. Other Grants and Activities

7.1. Regional Initiatives

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

[3] http://www.microbemagazine.org/index.php/10-2010-journal-highlights

- The project on *non-ribosomal 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 student (Aude Darracq) is co-supervised on this subject.
- 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
- Since 2008, the team leads a monthly local working group on parallel computation on GPU. This group gathers people from 5 different teams in LIFL, INRIA Lille, and Université Lille 1, and obtained financial support (BQR) from Université Lille 1 for equipment.

7.2. National Initiatives

- ANR CoCoGen (2008-2011). The goal of this project is to study new methods for comparison of complete genomes. The project is coordinated by E. Rivals (LIRMM, Montpellier). Others participants are MIG and UBLO teams of INRA (Jouy-en-Josas), INA-PG (Paris). The budget of this project is managed by the Montpellier partner. It covers travel fees to attend meetings.
- 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 SEQUOIA2. The topic is *Nouvelles approches algorithmiques et bioinformatiques pour l'analyse des grandes masses de données issues des séquenceurs de nouvelle génération*.
- NCRNA, RNG-Renabi, national network for bioinformatics (2007-2009). The objective is to develop an open-source annotation platform for noncoding RNA genes (see RNAspace in Section 6.3). This project involves the bioinformatics platforms of Génopole Toulouse-Midi-Pyrénées and SEQUOIA, and is supervised by C. Gaspin (Toulouse-Midi-Pyrénées). New support is planned for 2011.
- Work on mapping SOLiD reads: Group of E. Barillot (Institut Curie, Paris)
- The following scientists were invited in the past year to give a talk at the team seminar: Cédric Saule (LRI, Université Paris-Sud), Mickaël Salson (LITIS, Université Rouen), Giulia Chinetti (Unité Inserm U1011), Philippe Lefebvre (Unité Inserm U1011), Dominique Lavenier (INRIA Rennes), Matthieu Raffinot (LIAFA, Université Paris Diderot).

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. Following the 1-month visit of Peter Steffen, from University Bielefeld (Germany), M. Giraud went one month in Bielefeld (april 2010), developing a first OpenCL prototype implementation.

8. Dissemination

8.1. Animation of the scientific community

- 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.
- 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, ...
- We organized a joint scientific meeting with Institut Pasteur de Lille and Université Lille 2 on models for integration of heterogeneous complex biological data in May 2010 (55 participants).
- Participation to the scientific committee of the national ARENA group (bioinformatics of noncoding RNAs) created in 2004. This groups organizes a national workshop on RNA bioinformatics every year.

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 protéomique, master biologie-santé, master génie cellulaire et moléculaire, master interface physique-chimie*) in an Engineering School (Polytech'Lille), as well as in permanent education (for researchers, engineers and technicians).

8.2.1. Lectures on bioinformatics, University of Lille 1

- Organization of a lecture series on *Algorithms and computational biology*, master in computer science (M2), 15h (L. Noé, M. Pupin, H. Touzet)
- *Bioinformatics*, master génomique et protéomique (M1), 54h (L. Noé)
- *Bioinformatics*, master génomique et microbiologie (M1), 24h (A. Ouangraoua)
- *Bioinformatics*, master protéomique (M2), 24h (M. Giraud)
- *Bioinformatics*, master génie cellulaire et moléculaire (M2), 15h (J.-S. Varré)
- *Bioinformatics*, master biologie-santé (M2), 14h (H Touzet)
- *Algorithms in bioinformatics*, master in computer sciences (M1), 21h (H. Touzet)

8.2.2. Teaching in computer science, University of Lille 1

- *Programming (Pascal)*, second year of bachelor, 36h (J.-S. Varré)
- *Algorithms (Ada)*, third year of bachelor, 58h (L. Noé)
- *Networks*, third year of bachelor, 36h (L. Noé)
- *System*, third year of bachelor, 36h (L. Noé)
- *Algorithms and Data Structures*, third year of bachelor, 60h (J.-S. Varré)
- *Software project*, third year of bachelor, 35h (J.-S. Varré)
- *Algorithmics*, second year of bachelor, 30h (A. Saffarian)

8.3. Popular science

- An article describing the algorithm of structural pattern search was published at the french `)i(nterstice` popular science website [4] (M. Pupin)
- We organized an two-months exhibition of bioinformatics puzzles at Palais de la découverte (science museum in Paris), in the "Un chercheur, un manip" series[5]. These puzzles explain the basics of sequence assembly, RNA secondary structures and phylogenetic reconstruction. We were present at the exhibition all week-ends and school holidays in this period (M. Giraud, L. Noé, A. Ouangraoua, M. Pupin, A. Saffarian, M. Salson, and two collegues of the Symbiose team in Rennes)

[4] Des peptides à explorer, http://interstices.info/jcms/c_42760/des-peptides-a-explorer

[5] Le génome, un giga-puzzle informatique, http://www.lifl.fr/~giraud/puzzles

8.4. Administrative activities

- H. Touzet is a national representative (*chargée de mission*) for the Institute for Computer Sciences (INS2I) in CNRS[6]. She is more specifically in charge of relationships between the Institute and biology sciences.
- Member of the INRIA evaluation commitee (M. Giraud)
- Scientific secretary of the Gilles Kahn PhD award commitee (M. Giraud)
- Coordinator of ReNaBi-NE regional center (Réseaux National des Plates-formes de Bio-informatique Nord-Est) gathering the bioinformatics platforms of Lille (CIB), Strasbourg (IGBMC), Vandoeuvre-lès-Nancy (SIDR, MBI) and Reims (MMP). This center is one partner of ReNaBi-IFB project (French Bioinformatics Institute) (M. Pupin)
- Member of ITMO Genetics, Genomics and Bioinformatics, alliance AVIESAN (H .Touzet)
- Reviewer for ANR Jeunes Chercheurs program (J.-S. Varré)
- Reviewer for CIR-JEI (Crédit Impôt Recherche - Jeune Entreprise Innonvante) (J.-S. Varré)
- Reviewer for Ecole Doctorale de Bretagne (J.-S. Varré)
- Member of the reviewing committee for INRIA associate teams (A. Ouangraoua)
- Reviewer for PEPS and PICS programs (H.Touzet)
- Head of PPF bioinformatics – University Lille 1 (H. Touzet)
- Head of CIB, Lille bioinformatics platform (M. Pupin)
- Member of hiring committee *(jury d'audition)* of INRIA Rennes (M. Giraud)
- Members of hiring committee *(Commission des Spécialistes)* of the University Lille 1 (H. Touzet and M. Pupin), Université Lille 2 (H. Touzet), Univesrité Bordraux 1 (H. Touzet), Université de Nice (M. Pupin)
- Coordinator for the RAweb 2010 of INRIA Lille Nord Europe (A. Ouangraoua)
- Member of Cordi/postdoc INRIA Lille commission (M. Pupin)
- Member of the Men/Women egality working group, University Lille 1 (M. Pupin)
- Member of the LIFL Laboratory council (H. Touzet)
- Member of the INRIA Lille center commitee (J.-S. Varré)

8.5. PhD theses committess

- Member of the thesis committee of Arthur Tapi, Lille 1 University (M. Pupin)
- Member of the thesis committee of Catherine Eng, Nancy University (J.-S. Varré)
- Member of the thesis committee of Y.P. Deniélou, Université Joseph Fourier, Anthony Mathelier, Université Paris 6, Antonin Marchais, Université Paris 11, habilitation of V. Ranwez, Université Montpellier 2 (H. Touzet)

8.6. Editorial and reviewing activities

- Program committee of ICCS/WEPA 2010 (M. Giraud), JOBIM 2010 (H. Touzet), Recomb-CG 2010 (A. Ouangraoua),
- Reviewer for the journals Parallel Computing (M. Giraud), JDA (A. Ouangraoua), BMC Bioinformatics (H. Touzet)
- Reviewer for the conferences CARI 2010 (A. Ouangraoua), FUN 2010 (M. Giraud), JOBIM 2010 (M. Giraud, L. Noé), SAC 2011 (J.-S. Varré, L. Noé, H. Touzet), STACS 2011 (M. Giraud),

[6] CNRS: National Center for Scientific Research

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] A. DARRACQ. *Evolution des génomes mitochondriaux de plantes - Approche de génomique comparative chez Zea mays et Beta vulgaris*, Université Lille 1, 2010.

Articles in International Peer-Reviewed Journal

[2] S. CABOCHE, V. LECLÈRE, M. PUPIN, G. KUCHEROV, P. JACQUES. *Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity*, in "Journal of bacteriology", October 2010, vol. 192, n^o 19, p. 5143-5150 [*DOI :* 10.1128/JB.00315-10], http://jb.asm.org/cgi/content/abstract/192/19/5143.

[3] C. CHAUVE, H. GAVRANOVIC, A. OUANGRAOUA, E. TANNIER. *Yeast ancestral genome reconstructions: the possibilities of computational methods*, in "Journal of Computational Biology", September 2010, vol. 2010, p. 1097-1112, http://www.liebertonline.com/doi/full/10.1089/cmb.2010.0092.

[4] A. DARRACQ, J.-S. VARRÉ, P. TOUZET. *A scenario of mitochondrial genome evolution in maize based on rearrangement events*, in "BMC Genomics", April 2010, vol. 11, n^o 233 [*DOI :* 10.1186/1471-2164-11-233], http://www.biomedcentral.com/1471-2164/11/233.

[5] M. GIRAUD, J.-S. VARRÉ. *Parallel Position Weight Matrices Algorithms*, in "Parallel Computing", 2010, to appear, http://dx.doi.org/10.1016/j.parco.2010.10.001.

[6] M. GÎRDEA, G. KUCHEROV, L. NOÉ. *Back-translation for discovering distant protein homologies in the presence of frameshift mutations*, in "Algorithms for Molecular Biology", January 2010, vol. 5, n^o 6 [*DOI :* 10.1186/1748-7188-5-6], http://www.almob.org/content/5/1/6.

[7] 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.html.

[8] A. OUANGRAOUA, A. BERGERON. *Combinatorial structure of genome rearrangements scenarios*, in "Journal of Computational Biology", September 2010, vol. 2010, p. 1129-1144, http://www.liebertonline.com/doi/full/10.1089/cmb.2010.0126.

[9] 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", 2010, to appear.

[10] L. TONON, H. TOUZET, J.-S. VARRÉ. *TFM-Explorer: mining cis-regulatory regions in genomes*, in "Nucleic Acids Research", June 2010, vol. 38, n^o suppl_2, p. W286-292 [*DOI :* 10.1093/NAR/GKQ473], http://nar.oxfordjournals.org/cgi/content/abstract/38/suppl_2/W286.

International Peer-Reviewed Conference/Proceedings

[11] H. BANNAI, M. GIRAUD, K. KUSANO, W. MATSUBARA, A. SHINOHARA, J. SIMPSON. *The Number of Runs in a Ternary Word*, in "Prague Stringology Conference 2010 (PSC 2010)", 2010, http://www.stringology.org/event/2010/p16.html.

[12] 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)", 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/.

[13] A. OUANGRAOUA, A. BERGERON, K. SWENSON. *Ultra-perfect Sorting Scenarios*, in "RECOMB-Comparative Genomics, LNBI 6398", E. TANNIER (editor), Lecture Notes in Bioinformatics, October 2010, vol. 6398, p. 50-61.

[14] A. OUANGRAOUA, K. SWENSON, C. CHAUVE. *An approximation algorithm for computing a parsimonious first speciation in the gene duplication model*, in "RECOMB-Comparative Genomics, LNBI 6398", E. TANNIER (editor), Lecture Notes in Bioinformatics, October 2010, vol. 6398, p. 290-302.

Scientific Books (or Scientific Book chapters)

[15] N. PISANTI, M. GIRAUD, P. PETERLONGO. *Filters and Seeds Approaches for Fast Homology Searches in Large Datasets*, in "Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications", M. ELLOUMI, A. Y. ZOMAYA (editors), Wiley, 2010, (to appear), http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470505192.html.

[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", World Scientific, 2010, to appear, http://www.worldscibooks.com/lifesci/7972.html.

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 [*DOI :* 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].

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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 [*DOI :* 10.1109/TCBB.2009.4], http://www.lifl.fr/~noe/files/pp_TCBB09_preprint.pdf.

[33] 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.

[34] M. Salson, T. Lecroq, M. Léonard, L. Mouchard. *Dynamic Extended Suffix Array*, in "Journal of Discrete Algorithms", 2010, vol. 8, p. 241–257.

[35] P. Steffen, R. Giegerich, M. Giraud. *GPU Parallelization of Algebraic Dynamic Programming*, in "Parallel Processing and Applied Mathematics / Parallel Biocomputing Conference (PPAM / PBC 09)", Lecture Notes in Computer Science (LNCS), 2009, vol. 6068, p. 290-299.

[36] 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.