UMR IRISA

Activity report 2013

Dpt 7: **DATA AND KNOWLEDGE MANAGEMENT**

Team GENSCALE

Scalable, Optimized and Parallel Algorithms for Genomics

Rennes

Table of contents

Project-Team GENSCALE

Keywords: Computational Biology, Next Generation Sequencing, Genomics, Protein Structure, Big Data, Parallelism

Creation of the Team: 2012 January 01*, updated into Project-Team:* 2013 January 01.

1. Members

Research Scientists

Dominique Lavenier [Team leader, CNRS, Senior Researcher, HdR]
Claire Lemaitre [Inria, Researcher]
Pierre Peterlongo [Inria, Researcher]

Faculty Members

Rumen Andonov [Univ. Rennes I, Professor, HdR]
Antonio Mucherino [Univ. Rennes I, Associate Professor]

Engineers

Susete Alves Carvalho [INRA]
Alexan Andrieux [Inria ADT Mapsembler]
Erwan Drezen [Inria, ANR GATB project]
Fabrice Legeai [INRA]
Lucas Galton [Inria, KALRAY, from May 2013 until Aug 2013]
Anaïs Gouin [INRA]
Charles Deltel [Inria, Research engineer, 50% time dedicated to the GenScale project]

PhD Students

Mathilde Le Boudic-Jamin [Univ. Rennes I]
Nicolas Maillet [Inria, ANR MAPPI project]
Guillaume Chapuis [ENS Cachan, Inria ANR MAPPI project]
François Moreews [INRA]
Erwann Scaon [Univ. Rennes I, until Jul 2013]

Post-Doctoral Fellows

Liviu Ciortuz [Inria, Conseil Régional de Bretagne, until Sep 2013]
Douglas Goncalves [CNRS, from Apr 2013]

Visiting Scientist

Van Hoa Nguyen [CNRS, until Feb 2013]

Administrative Assistant

Marie-Noëlle Georgeault [Inria]

Other

Guillaume Rizk [Inria, Research engineer, GATB Project, from May 2013]

2. Overall Objectives

2.1. High throughput processing of genomic data

GenScale is a bioinformatics research team. It focuses on methodological research at the interface between computer science and genomic. The main objective of the group is the design of scalable, optimized and parallel algorithms for processing the huge amount of genomic data generated by the recent advances of biotechnologies.

GenScale research activities cover the following domains:

- Next Generation Sequencing (NGS)
  - Fast and low memory footprint assembly
  - Variant extraction on raw data (without assembly)
  - Mapping
- High throughput sequence analysis
  - Bank to bank comparison
  - De novo comparative metagenomic
- 3D Protein structures
  - Alignment, comparison, classification
  - Conformation extraction from NMR data
- Bioinformatics workflow
  - Graphical capture
  - Parallel processing (cluster, cloud)

This pure computer science activity is maintained with strong collaboration with life science research groups on challenging genomic projects.

2.2. Highlights of the Year

- Creation of KoriScale, an Inria Innovation Laboratory (I-LAB) to promote technology transfers between GenScale and the Korilog Company. The research thematic is focusing on intensive genomic sequence comparison. It covers innovative string algorithm aspects together with multi-level parallelism implementation. [Letter in Emergences]
- For the 2nd consecutive year the GenScale team won the best poster award from the annual Jobim conference. We demonstrate the efficiency of our low memory footprint NGS assembly tools to assemble the *C. Elegans* genome on the Raspberry Pi board, a very low cost computer (< $ 50) equipped with limited memory resources (512 MB). [42] [Letter in Emergences]

3. Research Program

3.1. Introduction

To tackle challenges brought by the processing of huge amount of genomic data, the main strategy of GenScale is to merge the following computer science expertise:

- Data structure;
- Combinatorial optimization;
- Parallelism.

3.2. Data structure

To face the genomic data tsunami, the design of efficient algorithms involves the optimization of memory footprints. A key point is the design of innovative data structures to represent large genomic datasets into computer memories. Today's limitations come from their size, their construction time, or their centralized (sequential) access. Random accesses to large data structures poorly exploit the sophisticated processor cache memory system. New data structures including compression techniques, probabilistic filters, approximate string matching, or techniques to improve spatial/temporal memory access are developed [3].

3.3. Combinatorial optimization

For wide genome analysis, Next Generation Sequencing (NGS) data processing or protein structure applications, the main issue concerns the exploration of sets of data by time-consuming algorithms, with the aim of identifying solutions that are optimal in a predefined sense. In this context, speeding up such algorithms requires acting on many directions: (1) optimizing the search with efficient heuristics and advanced combinatorial optimization techniques [2], [5] or (2) targeting biological sub-problems to reduce the search space [7], [9]. Designing algorithms with adapted heuristics, and able to scale from protein (a few hundreds of amino acids) to full genome (millions to billions of nucleotides) is one of the competitive challenges addressed in the GenScale project.

3.4. Parallelism

The traditional parallelization approach, which consists in moving from a sequential to a parallel code, must be transformed into a direct design and implementation of high performance parallel software. All levels of parallelism (vector instructions, multi-cores, many-cores, clusters, grid, clouds) need to be exploited in order to extract the maximum computing power from current hardware resources [6], [8], [1]. An important specificity of GenScale is to systematically adopt a design approach where all levels of parallelism are potentially considered.

4. Application Domains

4.1. Sequence comparison

Historically, sequence comparison has been one of the most important topics in bioinformatics. BLAST is a famous software tool particularly designed for solving problems related to sequence comparisons. Initially conceived to perform searches in databases, it has mostly been used as a general-purpose sequence comparison tool. Nowadays, together with the inflation of genomic data, other software comparison tools that are able to provide better quality solutions (w.r.t the ones provided by BLAST) have been developed. They generally target specific comparison demands, such as read mapping, bank-to-bank comparison, meta-genomic sample analysis, etc. Today, sequence comparison algorithms must clearly be revisited to scale up with the very large number of sequence objects that new NGS problems have to handle.

4.2. Genome comparison

This application domain aims at providing a global relationship between genomes. The problem lies in the different structures that genomes can have: segments of genome can be rearranged, duplicated or deleted (the alignment can no longer be done in one piece). Therefore one major aim is the study of chromosomal rearrangements, breaking points, structural variation between individuals of the same species, etc. However, even analyses focused on smaller variations such as Single Nucleotide Polymorphisms (SNP) at the whole genome scale are different from the sequence comparison problem, since one needs first to identify common (orthologous) parts between whole genome sequences and thus obtain this global relationship (or map) between genomes. New challenges in genome comparison are emerging with the evolution of sequencing techniques. Nowadays, they allow for comparing genomes at intra-species level, and to deal simultaneously with hundreds or thousands of complete genomes. New methods are needed to find the sequence and structural variants between such a large number of non-assembled genomes. Even for the comparison of more distant species, classical methods must be revisited to deal with the increasing number of genomes but more importantly their decreasing quality: genomes are no longer fully assembled nor annotated.

4.3. Protein comparison

Comparing protein is important for understanding their evolutionary relationships and for predicting their structures and their functions. While annotating functions for new proteins, such as those solved in structural genomics projects, protein structural alignment methods may be able to identify functionally related proteins when the sequence identity between a given query protein and the related proteins are low (i.e. lower than 20%). Moreover, protein comparison allows for solving the so-called protein family identification problem. Given an unclassified protein structure (query), the comparison of protein structures can be used for assigning a score measuring the "similarity" between the query and the proteins belonging to a set of families. Based on this score, the query is assigned to one of the families of the set. The knowledge acquired by performing such analyses can then be exploited in methods for protein structure prediction that are based on a homology modeling approach.

5. Software and Platforms

5.1. Next Generation Sequencing

Participants: Alexan Andrieux, Dominique Lavenier, Claire Lemaitre, Nicolas Maillet, Pierre Peterlongo, Guillaume Rizk, Erwan Drezen, Charles Deltel.

- **Genome assembly** [contact: P. Peterlongo]
  - **Minia : ultra low memory footprint assembly** Minia is a short-read assembler based on a de Bruijn graph, capable of assembling a human genome on a desktop computer in a day. The output of Minia is a set of contigs. Minia produces results of similar contiguity and accuracy to other de Bruijn assemblers (e.g. Velvet). http://minia.genouest.org/
  - **Mapsembler: targeted assembly software.** Mapsembler is a targeted assembly software. From sets of NGS raw reads and a set of input sequences (starters), it determines if each starter could be constructed from the reads. Then for each "read-coherent" starter, Mapsembler outputs its sequence neighborhood as a linear sequence or as a graph, depending on the user choice. http://colibread.inria.fr/mapsembler2/
  - **Bloocoo: memory-efficient read correction** Bloocoo is a software to identify sequencing errors in short-read datasets and correct them. It is based on an efficient data structure that enables to keep a very low memory footprint. http://gatb.inria.fr
- **Variant detection** [contact: C. Lemaitre]
  - **discoSnp and kisSplice : variant identification without the use of a reference genome**. discoSnp is a tool to find single nucleotide polymorphisms (SNP) by comparing two sets of raw NGS reads. http://colibread.inria.fr/discosnp/ KisSplice finds alternative splicings but also short insertions, deletions and duplications, SNPs and sequencing errors in one or two RNA-seq sets, without assembly nor mapping on a reference genome. http://colibread.inria.fr/software/kissplice/
  - **Kissreads: quantification of variants** Kissreads considers sets of NGS raw reads and a set of input sequences (starters). Mapping reads to each starter, it provides quantitative (coverage depth) and qualitative (mapped read quality) information about each starter.
  - **MindTheGap : detection of large insertions** MindTheGap is a tool to detect large insertion events in re-sequencing data with respect to a reference genome. http://gatb.inria.fr
- **Read mapping** [contact: D. Lavenier]
  - **GASSST: short reads mapper** The GASSST software (Global Alignment Short Sequence Search Tool) is a general purpose mapper. GASSST finds global alignments of short DNA sequences against large DNA banks. One main characteristic of GASSST is its ability to perform fast gapped alignments and to process long reads compared to other current similar tools. http://www.irisa.fr/symbiose/projects/gassst/

5.2. High throughput sequence analysis

Participants: Erwan Drezen, Dominique Lavenier, Claire Lemaitre, Nicolas Maillet, Pierre Peterlongo.

- **PLAST : efficient bank-to-bank alignments** PLAST (Parallel Local Alignment Search Tool) is a parallel alignment search tool for comparing large protein banks. PLAST runs 3 to 5 times faster than the NCBI-BLAST software. An improved version is commercialized by the Korilog Company, including the DNA bank-to-bank option. [contact: D. Lavenier] http://www.irisa.fr/symbiose/projects/plast/
- **Compareads : efficient comparison of large metagenomics NGS datasets** This software extracts similar DNA sequences (reads) between two metagenomic datasets. It requires a small and fixed amount of memory and can thus be used on huge datasets. [contact: P. Peterlongo] http://alcovna.genouest.org/compareads/

5.3. 3D Protein structures

Participants: Rumen Andonov, Guillaume Chapuis, Mathilde Le Boudic-Jamin, Antonio Mucherino.

- **CSA and DALIX** CSA (Comparative Structural Alignment) is a webserver for computing and comparing protein structure alignments. CSA is able to compute score-optimal alignments with respect to various inter-residue distance-based scoring schemes. [contact: R. Andonov] http://csa.project.cwi.nl/
- **A_purva** A_purva is a Contact Map Overlap maximization (CMO) solver. Given two protein structures represented by two contact maps, A_purva computes the amino-acid alignment which maximize the number of common contacts. [contact: R. Andonov] http://mobyle.genouest.org/cgi-bin/Mobyle/portal.py?forms::A_Purva
- **MD-Jeep** MD-jeep is a software tool for solving distance geometry problems. It is able to solve a subclass of instances of the problem for which a discrete reformulation can be supplied. We refer to this subclass of instances as the Discretizable Molecular Distance Geometry Problem (DMDGP). We employ a Branch & Prune (BP) algorithm for the solution of DMDGPs. [contact: A. Mucherino] http://www.antoniomucherino.it/en/mdjeep.php

5.4. HPC and Parallelism

Participants: Guillaume Chapuis, Dominique Lavenier, François Moreews, Charles Deltel.

- **QTLmap** QTLMap is a tool dedicated to the detection of Quantitative Trait Loci (QTL) from experimental designs in outbred population. QTLMap was recently ported to GPU and offers reduced run times. [contact: D. Lavenier] http://www.inra.fr/qtlmap/
- **SLICEE** (Service Layer for Intensive Computation Execution Environment) is part of the BioWIC project. This software proposes (1) to abstract the calls to the cluster scheduler by handling command submission; (2) to take care of exploiting the data parallelism with data specific methods; (3) to manage data using a cache references mecanism and route data between tasks. [contact: F. Moreews] http://vapor.gforge.inria.fr/

6. New Results

6.1. NGS methodology

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Anaïs Gouin, Fabrice Legeai.

- **Efficient Kmer counting**: Counting all the substrings of length k (k-mers) in DNA/RNA sequencing reads is the preliminary step of many bioinformatics applications. However, state of the art k-mer counting methods require that a large data structure resides in memory. Such structure typically grows with the number of distinct k-mers to count. We have developed a new streaming algorithm for that purpose which only requires a fixed user-defined amount of memory and disk space. This approach realizes a memory, time and disk trade-off. DSK is the first approach that is able to count all the 27-mers of a human genome dataset using only 4.0 GB of memory and moderate disk space (160 GB), in 17.9 h. DSK can replace a popular k-mer counting software (Jellyfish) on small-memory servers. [24]
- **Questionning the classical re-sequencing analyses approach**: Classical re-sequencing analyses are based on a first step of read mapping, then only mapped reads are taken into account in following analyses such as variant calling. We investigated the sources of unmapped reads in aphid re-sequencing data of 33 individuals, and we demonstrated that these reads contain valuable information that should not be discarded as usually done in such analyses. We proposed also an approach to extract this information, based on assembly and re-mapping. [34]
- **Repeat detection** A new algorithm was developed for detecting long similar fragments occurring at least twice in a set of biological sequences. The problem becomes computationally challenging when the frequency of a repeat is allowed to increase and when a non-negligible number of insertions, deletions and substitutions are allowed. The proposed algorithm, called Rime (for Repeat Identification: long, Multiple, and with Edits) performs this task, and manages instances whose size and combination of parameters cannot be handled by other currently existing methods. To the best of our knowledge, Rime is the first algorithm that can accurately deal with very long repeats (up to a few thousands), occurring possibly several times, and with a rate of differences (substitutions and indels) allowed among copies of a same repeat of 10-15% or even more. [17]

6.2. NGS applications

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Fabrice Legeai.

- **Participation to an international competition of assembly**: The process of generating raw genome sequence data continues to become cheaper, faster, and more accurate. However, assembly of such data into high-quality finished genome sequences remains challenging. Many genome assembly tools are available, but they differ greatly in terms of their performance and in their final output. More importantly, it remains largely unclear how to best assess the quality of assembled genome sequences. In this context, we have participated to the Assemblathon-2 competitions, which purpose was to assess current state-of-the-art methods in genome assembly. Globally, the cumulative z-scores of different assembly criteria set our assembly strategy in the 4th position compared to other competitors (21 groups). [12]
- **Assembly on Raspberri Pi**: Current Assembly tools require computers with large memory configuration. In order to demonstrate the efficiency of our low memory footprint assembly tools, we assemble the genome of C. Elegans (100 Mbp) on the raspberry PI computer, a small system equipped with only 512 MB RAM and 32 GB flash drive. [42]
- **SNP detection on the tick** We took part of a population genetic study on the tick species *Ixodes ricinus*, the main vector species of human and animal vector-borne diseases in Europe. In this framework, we proposed the first identification of a set of SNPs isolated from the genome of *I. ricinus*, by applying, among others a new tool developed in the GenScale team: discoSnp. The main advantage of this tool is to be able to detect SNPs without the use of a reference genome, which is crucially lacking for the tick species. Among the detected SNPs, 384 were selected, according to their minimal and maximal coverage and context sequences for experimental validation. Among them, 368 (95.8%) were biologically validated, demonstrating the precision of discoSNP.[23]

- **NGS analyses on insect models** We achieved the transcriptome assembly and analyzed the differential expression of an important noctuid pest. [22], [18]. Using gene expression data (RNA-Seq) in males, sexual females and asexual females of the pea aphid, we confirm theoretical models suggesting that the evolution of sex-biased gene expression may restrict the product of a sexually antagonistic allele to the sex it benefits.[19]
- **Genome sequencing and annotation**: We participated in the sequencing and annotation of several bacterial species of the Mollicute group. These bacteria are important pathogens of ruminants. The sequencing and annotation of their genomes confirmed their pathogenic features and phylogenetic location in the tree of Mollicutes. This is the first step before comparative genome analyses to unravel the genetic basis of mycoplasma pathogenicity and host specificity. [15], [16], [21]

6.3. HPC and parallelism

Participants: Dominique Lavenier, Rumen Andonov, Guillaume Chapuis, François Moreews, Charles Deltel.

- **Improving time performances of Mapping quantitative trait loci (QTL)** : we have developed a fast implementation of QTLMap, which takes advantage of the data parallel nature of the problem by offsetting heavy computations to a graphics processing unit (GPU). This new implementation performs up to 75 times faster than the previous multicore implementation, while maintaining the same results and level of precision . This speedup allows one to perform more complex analyses, such as linkage disequilibrium linkage analyses (LDLA) and multiQTL analyses, in a reasonable time frame. [13]
- **Integration of parallelism in bioinformatics workflows**: We propose a Model-Driven Architecture approach for capturing the complete design process of bioinformatics workflows. This approach is applied to graphical workflow editors and allows to quickly convert a workflow prototype in a parallel implementation. This work can have an impact on the way bioinformaticians implement their analysis and increase their productivity.[30]
- **Parallel assembly on FPGAs**: This research work proposes a method to reduce the overall time for assembly by using pre-processing of the short read data on FPGAs and processing its output using Velvet. We demonstrate significant speed-ups with slight or no compromise on the quality of the assembled output.[32]
- **All-Pairs Shortest Paths with multi-GPU** We propose a new algorithm for the All-Pairs Shortest Paths problem for graphs with good partitioning properties and its multi-GPU implementation. Our implementation targets large graphs (up to 10^6 vertices) and allows graphs with negative edges to be computed. [35]

6.4. Protein structures

Participants: Rumen Andonov, Guillaume Chapuis, Dominique Lavenier, Mathilde Le Boudic-Jamin, Antonio Mucherino, Douglas Goncalves.

- **A book on distance geometry problems (DGP).** This is a collection of invited papers on the topic "distance geometry" [38]. Among the other contributions, it contains a survey on "distance geometry" and "structural biology", which tries to function as a bridge between two scientific communities: computer science and biology. It presents some recent developments in the field by using a language common to the two communities [37]. In another contribution, the complexity of the DGP is discussed: even if this problem is NP-hard in general, we noticed a polynomial complexity on instances of DGP related to protein conformations (in the case all the available distances are exact)[36].
- **DGP with interval data.** In our preliminary works on the discretization of the Distance Geometry Problem (DGP), we considered instances where all distances were supposed to be exactly known. When biological molecules are concerned, however, this is not generally the case. We worked therefore for considering the full-atom representation of the protein backbone, where some of

the distances are subject to uncertainty within a given nonnegative interval. We showed that the discretization is still possible in this case, and proposed the iBP algorithm to solve the discretized DGP. [20]

- **New pruning device for DGP.** After the discretization, DGPs can be solved by a branch-and-prune (BP) algorithm, which is potentially able to enumerate the entire solution set. This solution set, however, can be very large for some instances, while only the most energetically stable conformations are of interest. We worked therefore for integrating the BP algorithm with two new energy-based pruning devices. Our computational experiments showed that the newly added pruning devices were actually able to improve the performance of the algorithm, as well as the quality (in terms of energy) of the conformations in the solution set. [28]
- **Discretization orders for the DGP.** The main assumption that allows for the discretization of DGPs is strongly based on the order in which the atoms of the molecule are considered. The "natural" order of the atoms in the amino acid chain does not always allow for the discretization. We tried to find discretization orders in several ways, based on different approaches. In [31], we extended a previously proposed greedy algorithm that is able to deal with interval data (inexact distances). In [27], we handcrafted some discretization orders for the side chains of the amino acids involved in the protein synthesis. In [29], we proposed a heuristic, which outperforms, on large instances, the greedy algorithm previously proposed.
- **DGP with Clifford Algebra** The BP algorithm for the DGP is based on a search on the tree, where nodes of the tree belonging to a common layer provide the possible positions for the same atom of the molecule. When interval data are given, a curve in 3d (containing the possible positions for the atom) can be associated to one of such nodes. Since it is generally not necessary to have protein conformations with a precision higher than 1A, sample points on these curves can be chosen. The way to choose these sample points is not, however, a simple task. This is the reason why we are trying to make this selection process adaptive, by exploiting Clifford Algebra to this purpose. Preliminary studies in this direction were presented in [25]
- **Parallel seed-based approach to protein structure similarity detection** We have developed a new parallel heuristic-based approach to structural similarity detection between proteins that discovers multiple pairs of similar regions. We prove that returned alignments have RMSDc and RMSDd lower than a given threshold. Computational complexity is addressed by taking advantage of both fine- and coarse-grain parallelism. [26]
- **Datamining.** The selection of features that describe samples in sets of data is a typical problem in data mining. A crucial issue is to select a maximal set of pertinent features, because the scarce knowledge of the problem under study often leads to consider features which do not provide a good description of the corresponding samples. The concept of consistent biclustering of a set of data has been introduced to identify such a maximal set. The problem can be modeled as a 0–1 linear fractional program, which is NP-hard. We reformulated this optimization problem as a bilevel program, and we proposed a heuristic for its solution [39].

7. Bilateral Contracts and Grants with Industry

7.1. I-Lab Koriscale

In June 2013, GenScale and the Korilog Company created an Inria common structure research (I-LAB) called KoriScale. This is the outcome of a solid relationship, which has enable the transfer of the PLAST software (bank to bank genomic sequence comparison) from GenScale to Korilog. The resulting commercial product (Klast) is now 5 to 10 times faster than the reference software (Blast). The main research axe of the I-LAB focuses on comparing huge genomic and metagenomic datasets.

7.2. Sequence Comparison, Korilog

Intensive bank-to-bank comparison with Korilog : this collaborative project between the Korilog company and the GenScale team aims to investigate new research directions in the bank-to-bank sequence comparison problem. Two research axes are followed : constrained exploration of the search space and adaptation of the ORIS algorithm, developed by D. Lavenier for fast DNA comparison, to the protein sequences. It is funded for 3 months (Nov. 2012 - Feb. 2013).

7.3. Sequence Comparison, Kalray

Parallelization of PLAST on many cores : This collaboration aims to implement he PLAST software on the MPPA chip (256 cores) developed by the Kalray company. PLAST is a BLAST-like parallel implementation of the bank to bank genomic sequence comparison problem. More generally, the purpose, here, is to investigate the performances of the MPPA architecture on scientific life science software. This is a bilateral contract of 4 months, from April to August 2013.

7.4. Peapol

The Peapol project is funded by Sofiproteol company whose mission is to develop the French vegetable oil and protein industry, open up new markets, and ensure an equal distribution of value among its members. The Peapol project counts two collaborators, Biogemma, and INRA, the latter working in collaboration with the Genscale team, in charge of algorithmic research in the context of the project. This collaboration enabled to hire in the Genscale team Raluca Uricaru for 18 months on an INRA post doctoral position, followed by Susete Alves-Carvalho (engineer).

7.5. Rapsodyn

RAPSODYN is a long term project funded by the IA French program (Investissement d'Avenir) and several field seed companies, such as Biogemma, Limagrain and Euralis. The objective is the optimisation of the rapeseed oil content and yield under low nitrogen input. GenScale is involved in the bioinformatics workpackage, in collaboration with Biogemma's bioinformatics team, to elaborate advanced tools dedicated to polymorphism.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Program from Région Bretagne : MIRAGE

Participants: Liviu Ciortuz, Claire Lemaitre, Pierre Peterlongo.

The MIRAGE project is funded by Région Bretagne in the framework of the SAD call (Stratégie Attractivité Durable) which aims at attracting international post-doctorant for one year. The MIRAGE project was funded from Sept. 2012 until August 2013 and coordinated by C. Lemaitre. It enabled to hire Liviu Ciortuz as a postdoctoral student for 12 months, for developping new methods to detect complex variation (structural variations) in non-assembled NGS data.

8.1.2. Program from Région Bretagne : DGASP

Participants: Antonio Mucherino, Douglas Goncalves.

This project is funded by Région Bretagne in the framework of the SAD call (Stratégie Attractivité Durable), from April 2013 to March 2014 and coordinated by A. Mucherino. It enabled to hire Douglas Goncalves as a postdoctoral student for 12 months for working on a discretizable class of distance geometry problems. The project is in collaboration with Carlile Lavor (IMECC-UNICAMP, Brazil) and Jacques Nicolas (équipe Dyliss, IRISA).

8.1.3. Poly-BNF

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Erwann Scaon.

This projects aims to develop bioinformatics strategies for studying polyploid genomes. It is a one year project (09/2012 – 09/2013) funded by the University of Rennes 1. It is a joined project with CNRS/EcoBio lab and INRA/IGEPP lab in Rennes.

8.1.4. Partnership with IGDR

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Fabrice Legeai, Charles Deltel.

We collaborate with several teams of the Genomic and Development Institute of Rennes (IGDR) : Genetics of dog (detection of long non coding RNAs in collaboration with Thomas Derrien and Christophe Hitte) and Integrated Fonctional Genomics and Biomarkers (NGS analyses of glioblastoma cancer, project funded by INCa in collaboration with Marie de Tayrac and Jean Mosser).

8.1.5. Partnership with INRA

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Anaïs Gouin, Fabrice Legeai, François Moreews, Susete Alves Carvalho.

We have a strong and long term collaboration with biologists of INRA in Rennes : IGEPP and PEGASE units. This partnership concerns both service and research activities and is acted by the hosting of two ingineers (F. Legeai, F. Moreews) and by the co-supervision of two non permanent engineers (A. Gouin, S. Alves Carvalho). In particular, the collaboration with the IGEPP team includes several research projects in which Genscale is formally a partner : an INRA project PEAPOL including an industrial partner, Biogemma, and an ANR project SPECIAPHID. These projects fund the non-permanent INRA members.

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. MAPPI

Participants: Dominique Lavenier, Claire Lemaitre, Nicolas Maillet, Pierre Peterlongo.

The MAPPI project aims to develop new algorithms and Bioinformatics methods for processing high trougthput genomic data. It is funded by ANR call COSINUS and coordinated by M. Raffinot (LIAFA, Paris VII) from Oct 2010 to Dec. 2013.

8.2.1.2. FATINTEGER

Participants: Dominique Lavenier, François Moreews.

The FatInteger project aims to identify some of the transcriptional key players of animal lipid metabolism plasticity, combining high throughput data with statistical approaches, bioinformatics and phylogenetic. It is funded by ANR call BLANC and coordinated by F. Gondret from 2012 to 2015.

8.2.1.3. SPECIAPHID

Participants: Anaïs Gouin, Fabrice Legeai, Claire Lemaitre.

The SPECIAPHID project aims to understand the adaptation and speciation of pea aphids by re-sequencing and comparing the genomes of numerous aphid individuals. Genscale's task, as associate partner, is to apply and develop new methods to detect variation between re-sequenced genomes, and in particular complex variants such as structural ones. It is funded by ANR call BLANC and coordinated by J-C Simon (Inra, Rennes) from January 2012 to Dec. 2014.

8.2.1.4. ADA-SPODO

Participants: Rumen Andonov, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, François Moreews, Pierre Peterlongo.

The ADA-SPODO project aims at identifying all sources of genetic variation between two strains of an insect pest : Lepidoptera Spodoptera frugiperda in order to correlate them with host-plant adaptation and speciation. Genscale's task is to develop new efficient methods to compare complete genomes along with their post-genomic and regulatory data. It is funded by ANR call BLANC and coordinated by E. d'Alençon (Inra, Montpellier) from October 2012 to Dec. 2015.

8.2.1.5. RAPSODYN

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Erwann Scaon.

RAPSODYN is a long term project funded by the IA French program (Investissement d'Avenir) for 7.5 years (07/2012-12/2019). The objective is the optimisation of the rapeseed oil content and yield under low nitrogen input. GenScale is involved in the bioinformatics workpackage to elaborate advanced tools dedicated to polymorphism.

8.2.1.6. COLIB'READ

Participants: Pierre Peterlongo, Claire Lemaitre, Dominique Lavenier, Fabrice Legeai, Guillaume Rizk.

The main goal of the Colib'Read project is to design new algorithms dedicated to the extraction of biological knowledge from raw data produced by High Throughput Sequencers (HTS). The project proposes an original way of extracting information from such data. Usually, a generic assembly (pre-treatment) is applied to the data, and then, in a second step, any information of interest is extracted. Our aim is to avoid this protocol that leads to a significant loss of information, or generates chimerical results because of the heuristics used in the assembly. Instead, the project will propose a set of innovative approaches for extracting information of biological interest from HTS data, with methods that bypass any costly and often inaccurate assembly phase, not requiring the availability of a reference genome. It is funded by ANR call BLANC and coordinated by P. Peterlongo from March 2013 to February 2016. https://colibread.inria.fr/

8.2.1.7. GATB

Participants: Dominique Lavenier, Erwan Drezen, Pierre Peterlongo, Claire Lemaitre, Guillaume Rizk.

GATB (Genome Assembly Tool Box) is a project that aims to provide algorithms and tools for genome assembly. The strength of these algorithms comes from the underlying structure that has a low memory footprint, which enables to assemble genomes on a simple desktop computer. The GATB project will provide several software components, such as low level libraries, binaries and pipelines providing a full spectrum of tools for genome assembly. It is a 2 years ANR project started in February 2013. http://gatb.inria.fr

8.2.2. Programs from research institutions

8.2.2.1. Mapsembler

Participants: Alexan Andrieux, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

The Mapsembler project aims at finalizing and to distributing the Mapsembler tool. It is funded by Inria ADT call (2012) and coordinated by P. Peterlongo from oct. 2012 to sept. 2014. http://alcovna.genouest.org/mapsembler/

8.2.2.2. Mastodons

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

This project, funded by the CNRS Big Data program in 2012 and 2013, aims do investigate the challenge brought by the processing of high throughput sequencing genomic data. It is coordinated by D. Lavenier from June 2012 to December 2013.

8.2.2.3. Barcoding de nouvelle génération

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

This project is a join initiative between Genscale and LECA (Laboratoire d'Ecologie Alpine in Grenoble). It aims at developping new algorithmic approaches for the species identification from low coverage NGS data. It is funded by a PEPS program at CNRS/Inria and coordinated by C. Lemaitre from Sept. 2012 to Dec. 2013.

8.2.2.4. Structuring of NGS for diagnostic purpose in cancer

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

This 18 months project is funded by the national institute of cancer (INCa). Genscale is involved in the optimization of bioinformatics workflows to detect variants in glioblastoma cancer.

8.2.3. Cooperations

8.2.3.1. Inria Bamboo Team

Participants: Claire Lemaitre, Pierre Peterlongo.

We maintain a long term collaboration with Inria Bamboo Team on the problems of finding biological information, such as variants, in NGS raw data.

8.2.3.2. LIGM, Paris

Participant: Pierre Peterlongo.

P. Peterlongo collaborates with the LIGM lab in Paris (UMR 8049), on problems of large NGS raw data indexation.

8.2.3.3. LIX

Participant: Antonio Mucherino.

A. Mucherino collaborates since 5 years with LIX, Ecole Polytechnique, in Palaiseau on the distance geometry problem. We reformulated the problem as a combinatorial optimization problem and we conceived an ad-hoc algorithm for the solution of this class of problems.

8.3. European Initiatives

8.3.1. Collaborations with Major European Organizations

Partner: CWI, University of Amsterdam, (Netherland)

Subject of cooperation: Optimization algorithms for protein structures alignments.

8.4. International Initiatives

8.4.1. Inria International Partners

8.4.1.1. Informal International Partners

Partner: IMECC, UNICAMP, Campinas-SP (Brazil)

Subject: distance geometry, bioinformatics.

Partner:COPPE, Federal University of Rio de Janeiro (Brazil)

Subject: distance geometry, bioinformatics.

Partner: Los Alamos National Laboratory (lanl), Los Alamos (USA)

Subjects: Combinatorial algorithms (shortest paths, graph partitioning, combinatorial optimization) and algorithm engineering (efficient implementation of combinatorial algorithms)

8.5. International Research Visitors

8.5.1. Visits of International Scientists

- Van-Hoa Nguyen from University of Angiang, Viet Nam, visited GenScale for 3 months (nov. 2012 - feb. 2013). The visit was funded by the French Mastodons program from CNRS to research focusing on bioinformatics big data problem.

- Fatima Sapundzhi and Boyana Garkova, PhD students from South-West University, Neofit Rilski, Blagoevgrad (Bulgaria), visited the team for one month in October 2013. The visit was funded by the bulgarian ministery and focused on ligand-protein interaction structure problems in collaboration with R. Andonov and M. Le Boudic-Jamin.

8.5.2. Visits to International Teams

- R. Andonov has been invited by the Information Sciences Group (CCS-3) from Los Alamos National Laboratory (LANL) for one month (15 July - 15 August 2013).

9. Dissemination

9.1. Scientific Animation

9.1.1. Meeting organization and scientific animation

- **Seminar** A weekly seminar of bioinformatics is organized within the laboratory. Attendees are member of the ex-symbiose team (now teams Genscale, Dyliss and Genouest), biologists from Brittany and computer scientists from the laboratory. [web site: http://symbiose.irisa.fr/symbioseNextSeminars]
- **Conference** P. Peterlongo and D. Lavenier were organisation members of the Environmental Genomic Colloque, Rennes, 4,5,6 Nov. 2013. [web site:https://colloque.inra.fr/ge_rennes2013]

9.1.2. Conference program committees

- FPL'2013: 23nd International Conference on Field Programmable Logic and Applications [D. Lavenier]
- ICPADS'2013: 19th IEEE International Conference on Parallel and Distributed Systems [D. Lavenier]
- PBC'2013: Workshop on Parallel Computational Biology [D. Lavenier]
- Reconfig'2013: International Conference on ReConFigurable Computing [D. Lavenier]
- JOBIM'2013: French Colloqium in Biology, Mathematic anf Informatic [D. Lavenier, C. Lemaitre, P. Peterlongo]
- SeqBio 2013: Workshop on string algorithms [C. Lemaitre]
- DGA 2013: Workshop on Distance Geometry and Applications [R. Andonov, A. Mucherino]
- WCO 2013: 6th Workshop on Computational Optimization [R. Andonov, A. Mucherino]

9.1.3. Administrative functions: scientific committees, journal boards

- Member of the administrative council of ISTIC [R. Andonov]
- External evaluator for COST Action IC0805 on "Open European Network for High Performance Computing on Complex Environments" [R. Andonov]
- Member of evaluation committee AERES for LIGM [R. Andonov]
- Recruitment committees: 1 assistant professor [D. Lavenier], 1 professor [D. Lavenier]
- Member for PEPS-BMI program comittee (CNRS) [P. Peterlongo]
- Permanent expert for the MEI (International Expertise Mission), French Research Ministry [D. Lavenier]
- Member of the local Inria Rennes CDT (Technologic Transfer Commission) [D. Lavenier]
- Member of the scientific council of the INRA BIPAA Platform (BioInformatics Platform for Agroecosystems Arthropods) [D. Lavenier]

- Member of the scientific council of The GenOuest Platform (Bioinformatics Platform of BioGenOuest) [D. Lavenier]
- Member of the local Inria CORDIS comittee for PhD grants [C. Lemaitre]
- Representative of the environnemental axis of UMR IRISA [C. Lemaitre]
- Inria center referee of Scientific mediation [P. Peterlongo]
- Member of the redaction committee Ouest Inria [P. Peterlongo]
- publication reviewing for Bioinformatics, BMC Bioinformatics, BMC Research Notes, Briefings in Bioinformatics, Plos One, Journal of Computational Chemistry, Optimization Methods and Software [D. Lavenier, R. Andonov]

9.1.4. Invited talks

- A. Mucherino gave an invited talk at DGA13, Manaus, Amazonas, Brazil, June 24–27, 2013
- D. Lavenier gave an invited talk at Inria-INRA days, Sophia Antipolis, September 11-12 2013
- D. Lavenier gave an invited talk at ORAP Meeting, Scalay, October 10 2013
- D. Lavenier gave an invited talk at X-meeting, Recife, Brazil, November 3-7 2013
- P. Peterlongo gave an invited talk at Colloque "Détection, Gestion et Analyse du Polymorphisme des Génomes Végétaux" of INRA EPGV, Lusignan, April 8-10 2013.
- P. Peterlongo gave an invited talk at Workshop "Storage, Search and Annotation of Multiple Similar Genomes", Bielefeld, December 9-10 2013.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : A. Mucherino, R. Andonov, Graph algorithms, 90h, L3, Univ. Rennes 1, Rennes France.

Licence : C. Lemaitre, Statistics for biology, 10h, L3, Univ. Rennes 1, France

Licence: D. Lavenier, Architecture and System, 36h, MIT1, ENS Rennes

Licence : M. Le Boudic-Jamin, Preparation to C2I level 1 , 44h, L1, Univ. Rennes 1, France

Master : D. Lavenier, Intensive Computation of Genomic Data, 18h, ESEO, M1, Angers

Master : C. Lemaitre, P. Peterlongo, Text algorithmics for Bioinformatics, 46 h, M1, Univ. Rennes 1, France.

Master : C. Lemaitre, Dynamical systems for biological networks, 22h, M2, Univ. Rennes 1, France

Master : R. Andonov, A. Mucherino, Operations research, 95h, M1, Univ. Rennes 1, France.

Master : R. Andonov, Advanced algorithms, 15h, M1, Univ. Rennes 1, France.

Master : A. Mucherino, Initiation to systems and networks, 39h, M2, Univ. Rennes 1, France

Master : A. Mucherino, R. Andonov, P. Peterlongo, Sequence and structure algorithms, 36h, M2, Univ. Rennes 1, France

9.2.2. Supervision

PhD defense : Nicolas Maillet, *Comparaison de novo de données de séquençage issues de très grands échantillons métagénomiques»*[11], Univ. Rennes 1, defended on December 19th 2013, supervised by D. Lavenier and P. Peterlongo [online manuscript: http://tel.archives-ouvertes.fr/tel-00941922]

PhD defense : Guillaume Chapuis, *Exploiting parallel features of modern computer architectures in bioinformatics : Applications to genetics, structure comparison and large graph analyses*[10], Univ. Rennes 1, defended on December 18th 2013, supervised by D. Lavenier and R. Andonov [online manuscript: http://tel.archives-ouvertes.fr/tel-00912553]

PhD in progress : Mathilde Le Boudic-Jamin, *Structure et comparaison d'objets 3D: applications aux structures protéiques*, Univ. Rennes 1, started in October 2011, supervised by R. Andonov

PhD in progress : Erwan Scaon, *Modèles et algorithmes pour l'assemblage de novo de génomes à forte redondance*, Univ. Rennes 1, started in October 2012, supervised by D. Lavenier and C. Lemaitre

PhD in progress : François Moreews, *Environnement intégré de conception et d'exécution de workflows en bioinformatique: du prototypage au calcul intensif. Applications à la recherche de motifs de régulation dans les génomes*, Univ. Rennes 1, started in November 2012, supervised by D. Lavenier and S. Lagarigue

9.2.3. Juries

- *President of Ph-D thesis jury*. J. Lai, University of Rennes [D. Lavenier], O. Abdou-Arbi, University of Rennes 1 [R. Andonov]
- *Member of Ph-D thesis juries*. V. Silva Da Costa, Federal University of Rio de Janeiro (Brazil) [A. Mucherino]; R. Santos Alves, UNICAMP (Brazil) [A. Mucherino]; G. Chapuis, University of Rennes [D. Lavenier, R. Andonov]; Nicolas Maillet, University of Rennes [D. Lavenier, P. Peterlongo]; C. Yupeng, Nanyang Technological University, Singapor [D. Lavenier].
- *Member of Ph-D thesis comitees*. J. Boutte, University of Rennes [D. Lavenier]; A. Jeannin, University of Brest [D. Lavenier]; P. Nouhaud, University of Rennes [C. Lemaitre]; C. Mercier, University of Grenoble [C. Lemaitre]; S. Guizard, University of Tours [C. Lemaitre], A. Radulescu, university of Nantes [P. Peterlongo].

9.3. Popularization

- Participation to the event "A la découverte de la recherche" (presentation of the research activity to high school students) [C. Lemaitre, P. Peterlongo]
- Participation to the event "Professional Meeting" (talk: discovering bioinformatics), IUT Lannion [D. Lavenier]
- Production of a short movie of PhD subject popularization. [M. Le Boudic-Jamin] https://doctoriales2013.ueb.eu/content/posters

10. Bibliography

Major publications by the team in recent years

[1] R. ANDONOV, S. BALEV, N. YANEV. *Protein Threading: From Mathematical Models to Parallel Implementations*, in "INFORMS Journal on Computing", 2004, vol. 16, n° 4, pp. 393-405 [*DOI :* 10.1287/IJOC.1040.0092], http://joc.journal.informs.org/content/16/4/393.abstract

[2] R. ANDONOV, N. MALOD-DOGNIN, N. YANEV. *Maximum Contact Map Overlap Revisited*, in "Journal of Computational Biology", January 2011, vol. 18, n° 1, pp. 1-15 [*DOI :* 10.1089/CMB.2009.0196], http://hal.inria.fr/inria-00536624/en

[3] R. CHIKHI, G. RIZK. *Space-efficient and exact de Bruijn graph representation based on a Bloom filter*, in "Algorithms for Molecular Biology", 2013, vol. 8, n° 1, 22 p. [*DOI :* 10.1186/1748-7188-8-22], http://hal.inria.fr/hal-00868805

[4] F. LEGEAI, G. RIZK, T. WALSH, O. EDWARDS, K. GORDON, D. LAVENIER, N. LETERME, A. MEREAU, J. NICOLAS, D. TAGU, S. JAUBERT-POSSAMAI. *Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, Acyrthosiphon pisum*, in "BMC Genomics", 2010, vol. 11, n° 1, 281 p. [*DOI :* 10.1186/1471-2164-11-281], http://www.hal.inserm.fr/inserm-00482283

[5] A. MUCHERINO, C. LAVOR, L. LIBERTI, N. MACULAN. *The Discretizable Molecular Distance Geometry Problem*, in "Computational Optimization and Applications", 2012, vol. 52, pp. 115-146, http://hal.inria.fr/hal-00756940

[6] V. H. NGUYEN, D. LAVENIER. *PLAST: parallel local alignment search tool for database comparison*, in "Bmc Bioinformatics", October 2009, vol. 10, 329 p. , http://hal.inria.fr/inria-00425301

[7] P. PETERLONGO, R. CHIKHI. *Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer*, in "BMC Bioinformatics", March 2012, vol. 13, n° 48 [*DOI :* 10.1186/1471-2105-13-48], http://hal.inria.fr/hal-00675974

[8] G. RIZK, D. LAVENIER. *GASSST: Global Alignment Short Sequence Search Tool*, in "Bioinformatics", August 2010, vol. 26, n° 20, pp. 2534-2540, http://hal.archives-ouvertes.fr/hal-00531499

[9] G. A. T. SACOMOTO, J. KIELBASSA, R. CHIKHI, R. URICARU, P. ANTONIOU, M.-F. SAGOT, P. PETERLONGO, V. LACROIX. *KisSplice: de-novo calling alternative splicing events from RNA-seq data*, in "BMC Bioinformatics", March 2012, http://hal.inria.fr/hal-00681995

Publications of the year

Doctoral Dissertations and Habilitation Theses

[10] G. CHAPUIS. , *Exploiter les capacités parallèles des architectures modernes en bioinformatique applications à la génétique, la comparaison de structures et l'analyse de larges graphes*, École normale supérieure de Cachan - ENS Cachan, December 2013, http://hal.inria.fr/tel-00912553

[11] N. MAILLET. , *Comparaison de novo de données de séquençage issues de très grands échantillons métagénomiques : application sur le projet Tara Oceans*, Université Rennes 1, December 2013, http://hal.inria.fr/tel-00941922

Articles in International Peer-Reviewed Journals

[12] K. R. BRADNAM, J. N. FASS, A. ALEXANDROV, P. BARANAY, M. BECHNER, I. BIROL, S. BOISVERT, J. CHAPMAN, G. CHAPUIS, R. CHIKHI, H. CHITSAZ, W.-C. CHOU, J. CORBEIL, C. DEL FABBRO, T. RODERICK DOCKING, R. DURBIN, D. EARL, S. EMRICH, P. FEDOTOV, N. FONSECA, G. GANAPATHY, R. GIBBS, S. GNERRE, É. GODZARIDIS, S. GOLDSTEIN, M. HAIMEL, G. HALL, D. HAUSSLER, J. HIATT, I. HO, J. HOWARD, M. HUNT, S. JACKMAN, D. JAFFE, E. JARVIS, H. JIANG, S. KAZAKOV, P. KERSEY, J. KITZMAN, J. KNIGHT, S. KOREN, T.-W. LAM, D. LAVENIER, F. LAVIOLETTE, Y. LI, Z. LI, B. LIU, Y. LIU, R. LUO, I. MACCALLUM, M. MACMANES, N. MAILLET, S. MELNIKOV, D. NAQUIN, Z. NING, T. OTTO, B. PATEN, O. PAULO, A. PHILLIPPY, F. PINA-MARTINS, M. PLACE, D. PRZYBYLSKI, X. QIN, C. QU, F. RIBEIRO, S. RICHARDS, D. ROKHSAR, G. RUBY, S. SCALABRIN, M. SCHATZ, D. SCHWARTZ, A. SERGUSHICHEV, T. SHARPE, T. SHAW, J. SHENDURE, Y. SHI, J. SIMPSON, H. SONG, F. TSAREV, F. VEZZI, R. VICEDOMINI, B. VIEIRA, J. WANG, K. WORLEY, S. YIN, S.-M. YIU, J. YUAN, G. ZHANG, H. ZHANG, S. ZHOU, I. KORF1. *Assemblathon 2: evaluating de novo methods of genome assembly in three*

vertebrate species, in "GigaScience", July 2013, vol. 2, n^{o} 10 [*DOI :* 10.1186/2047-217X-2-10], http://hal.inria.fr/hal-00908747

[13] G. CHAPUIS, O. FILANGI, D. LAVENIER, J. M. ELSEN, P. LEROY. *Graphics Processing Unit-Accelerated Quantitative Trait Loci Detection*, in "Journal of Computational Biology", September 2013, vol. 20, n^{o} 9, pp. 672-686 [*DOI :* 10.1089/CMB.2012.0136], http://hal.inria.fr/hal-00903794

[14] R. CHIKHI, G. RIZK. *Space-efficient and exact de Bruijn graph representation based on a Bloom filter*, in "Algorithms for Molecular Biology", 2013, vol. 8, n^{o} 1, 22 p. [*DOI :* 10.1186/1748-7188-8-22], http://hal.inria.fr/hal-00868805

[15] E. DORDET-FRISONI, E. BARANOWSKI, A. BARRÉ, A. BLANCHARD, M. BRETON, C. COUTURE, V. DUPUY, P. GAURIVAUD, D. JACOB, C. LEMAITRE, L. MANSO-SILVÁN, M. NIKOLSKI, L.-X. NOUVEL, F. POUMARAT, P. SIRAND-PUGNET, P. THÉBAULT, S. THEIL, F. THIAUCOURT, C. CITTI, F. TARDY. *Draft Genome Sequences of Mycoplasma auris and Mycoplasma yeatsii, Two Species of the Ear Canal of Caprinae*, in "Genome Announcements", 2013, vol. 1, n^{o} 3 [*DOI :* 10.1128/GENOMEA.00280-13], http://hal.inria.fr/hal-00907449

[16] V. DUPUY, P. SIRAND-PUGNET, E. BARANOWSKI, A. BARRÉ, M. BRETON, C. COUTURE, E. DORDET-FRISONI, P. GAURIVAUD, D. JACOB, C. LEMAITRE, L. MANSO-SILVÁN, M. NIKOLSKI, L.-X. NOUVEL, F. POUMARAT, F. TARDY, P. THÉBAULT, S. THEIL, C. CITTI, A. BLANCHARD, F. THIAUCOURT. *Complete Genome Sequence of Mycoplasma putrefaciens Strain 9231, One of the Agents of Contagious Agalactia in Goats*, in "Genome Announcements", 2013, vol. 1, n^{o} 3 [*DOI :* 10.1128/GENOMEA.00354-13], http://hal.inria.fr/hal-00907450

[17] M. FEDERICO, P. PETERLONGO, N. PISANTI, M.-F. SAGOT. *Rime: Repeat identification*, in "Discrete Applied Mathematics", March 2013 [*DOI :* 10.1016/J.DAM.2013.02.016], http://hal.inria.fr/hal-00802023

[18] N. GLASER, A. GALLOT, F. LEGEAI, N. MONTAGNÉ, E. POIVET, M. HARRY, P.-A. CALATAYUD, E. JACQUIN-JOLY. *Candidate chemosensory genes in the Stemborer Sesamia nonagrioides*, in "Int J Biol Sci", 2013, vol. 9, n^{o} 5, pp. 481-95 [*DOI :* 10.7150/IJBS.6109], http://hal.inria.fr/hal-00916961

[19] J. JAQUIÉRY, C. RISPE, D. ROZE, F. LEGEAI, G. LE TRIONNAIRE, S. STOECKEL, L. MIEUZET, C. DA SILVA, J. POULAIN, N. PRUNIER-LETERME, B. SÉGURENS, D. TAGU, J.-C. SIMON. *Masculinization of the x chromosome in the pea aphid*, in "PLoS Genetics", August 2013, vol. 9, n^{o} 8 [*DOI :* 10.1371/JOURNAL.PGEN.1003690], http://hal.inria.fr/hal-00916967

[20] C. LAVOR, L. LIBERTI, A. MUCHERINO. *The interval Branch & Prune Algorithm for the Discretizable Molecular Distance Geometry Problem with Inexact Distances*, in "Journal of Global Optimization", 2013, vol. 56, n^{o} 3, pp. 855-871, http://hal.inria.fr/hal-00912660

[21] L. MANSO-SILVÁN, F. TARDY, E. BARANOWSKI, A. BARRÉ, A. BLANCHARD, M. BRETON, C. COUTURE, C. CITTI, E. DORDET-FRISONI, V. DUPUY, P. GAURIVAUD, D. JACOB, C. LEMAITRE, M. NIKOLSKI, L.-X. NOUVEL, F. POUMARAT, P. THÉBAULT, S. THEIL, F. THIAUCOURT, P. SIRAND-PUGNET. *Draft Genome Sequences of Mycoplasma alkalescens, Mycoplasma arginini, and Mycoplasma bovigenitalium, Three Species with Equivocal Pathogenic Status for Cattle*, in "Genome Announcements", 2013, vol. 1, n^{o} 3 [*DOI :* 10.1128/GENOMEA.00348-13], http://hal.inria.fr/hal-00907454

[22] E. POIVET, A. GALLOT, N. MONTAGNÉ, N. GLASER, F. LEGEAI, E. JACQUIN-JOLY. *A comparison of the olfactory gene repertoires of adults and larvae in the noctuid moth Spodoptera littoralis*, in "PLoS ONE", 2013, vol. 8, n^{o} 4 [*DOI :* 10.1371/JOURNAL.PONE.0060263], http://hal.inria.fr/hal-00916952

[23] E. QUILLERY, O. QUENEZ, P. PETERLONGO, O. PLANTARD. *Development of genomic resources for the tick Ixodes ricinus: isolation and characterization of Single Nucleotide Polymorphisms*, in "Molecular Ecology Ressources", October 2013, epub ahead of print [*DOI :* 10.1111/1755-0998.12179], http://hal.inria.fr/hal-00880072

[24] G. RIZK, D. LAVENIER, R. CHIKHI. *DSK: k-mer counting with very low memory usage*, in "Bioinformatics", January 2013, vol. 29, n^{o} 5, pp. 652-653, Bioinformatics journal requires that we post only the pre-print, which does not include modifications suggested by the reviewers [*DOI :* 10.1093/BIOINFORMATICS/BTT020], http://hal.inria.fr/hal-00778473

International Conferences with Proceedings

[25] R. ALVES, A. CASSIOLI, A. MUCHERINO, C. LAVOR, L. LIBERTI. *Adaptive Branching in iBP with Clifford Algebra*, in "Distance Geometry and Applications (DGA13)", Manaus, Amazonas, Brazil, 2013, pp. 65-69, http://hal.inria.fr/hal-00912698

[26] G. CHAPUIS, M. LE BOUDIC-JAMIN, R. ANDONOV, H. DJIDJEV, D. LAVENIER. *Parallel seed-based approach to protein structure similarity detection*, in "PPAM 2013", Varsovie, Poland, R. WYRZYKOWSKI (editor), Springer, May 2013, http://hal.inria.fr/hal-00881507

[27] V. COSTA, A. MUCHERINO, L. CARVALHO, N. MACULAN. *On the Discretization of iDMDGP instances regarding Protein Side Chains with Rings*, in "Distance Geometry and Applications (DGA13)", Manaus, Amazonas, Brazil, 2013, pp. 99-102, http://hal.inria.fr/hal-00912702

[28] D. GONÇALVES, A. MUCHERINO, C. LAVOR. *Energy-based Pruning Devices for the BP algorithm for Distance Geometry*, in "Workshop on Computational Optimization", Krakow, Poland, IEEE, 2013, pp. 335-340, http://hal.inria.fr/hal-00912670

[29] W. GRAMACHO, D. GONÇALVES, A. MUCHERINO, N. MACULAN. *A new Algorithm to Finding Discretizable Orderings for Distance Geometry*, in "Distance Geometry and Applications (DGA13)", Manaus, Amazonas, Brazil, 2013, pp. 149-152, http://hal.inria.fr/hal-00912706

[30] F. MOREEWS, D. LAVENIER. *Seamless Coarse Grained Parallelism Integration in Intensive Bioinformatics Workflows*, in "Proceedings of the 20th European MPI Users' Group Meeting", New York, NY, United States, EuroMPI '13, ACM, 2013, pp. 277–282 [*DOI :* 10.1145/2488551.2488588], http://hal.inria.fr/hal-00908842

[31] A. MUCHERINO. *On the Identification of Discretization Orders for Distance Geometry with Intervals*, in "Geometric Science of Information (GSI13)", Paris, France, Springer, 2013, pp. 231-238, http://hal.inria.fr/hal-00912677

[32] S. C. VARMA, P. KOLIN, M. BALAKRISHNAN, D. LAVENIER. *FAssem : FPGA based Acceleration of De Novo Genome Assembly*, in "Proceeding of The 21st Annual International IEEE Symposium on Field Programmable Custom Computing Machines", Seattle, United States, April 2013, FCCM'2013, http://hal.inria.fr/hal-00829055

National Conferences with Proceedings

[33] A. BRETAUDEAU, O. DAMERON, F. LEGEAI, Y. RAHBÉ. *AphidAtlas : avancées récentes*, in "Proceedings of BAPOA 2013 MOP. INRA, CIRAD", Montpellier, France, M. UZEST (editor), 2013, [In French], http://hal.inria.fr/hal-00913155

[34] A. GOUIN, P. NOUHAUD, F. LEGEAI, G. RIZK, J.-C. SIMON, C. LEMAITRE. *Whole genome re-sequencing : lessons from unmapped reads*, in "Journées Ouvertes Biologie Informatique Mathématiques", Toulouse, France, July 2013, http://hal.inria.fr/hal-00907446

Conferences without Proceedings

[35] G. CHAPUIS, H. DJIDJEV, R. ANDONOV, S. THULASIDASAN, D. LAVENIER. *Efficient Multi-GPU Algorithm for All-Pairs Shortest Paths*, in "IPDPS 2014", Phoenix, United States, Manish Parashar, May 2014, http://hal.inria.fr/hal-00905738

Scientific Books (or Scientific Book chapters)

[36] L. LIBERTI, C. LAVOR, A. MUCHERINO. *The Discretizable Molecular Distance Geometry Problem seems Easier on Proteins*, in "Distance Geometry: Theory, Methods and Applications", A. MUCHERINO, C. LAVOR, L. LIBERTI, N. MACULAN (editors), Springer, 2013, pp. 47-60, http://hal.inria.fr/hal-00912689

[37] T. E. MALLIAVIN, A. MUCHERINO, M. NILGES. *Distance Geometry in Structural Biology*, in "Distance Geometry: Theory, Methods and Applications", A. MUCHERINO, C. LAVOR, L. LIBERTI, N. MACULAN (editors), Springer, 2013, pp. 329-350, http://hal.inria.fr/hal-00912683

[38] A. MUCHERINO, C. LAVOR, L. LIBERTI, N. MACULAN. , *Distance Geometry: Theory, Methods and Applications*, Springer, 2013, 410 p. , http://hal.inria.fr/hal-00912679

[39] A. MUCHERINO, L. LIBERTI. *A VNS-based Heuristic for Feature Selection in Data Mining*, in "Hybrid Metaheuristics", E.-G. TALBI (editor), Springer, 2013, pp. 353-368, http://hal.inria.fr/hal-00912692

Research Reports

[40] M. C. DE COLA, A. MUCHERINO, G. FELICE, L. LIBERTI. , *A Branch-and-Prune Algorithm for the Sensor Network Localization Problem*, 2013, http://hal.inria.fr/hal-00912710

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

Other Publications

[42] G. COLLET, G. RIZK, R. CHIKHI, D. LAVENIER. *MINIA on a Raspberry Pi, Assembling a 100 Mbp Genome on a Credit Card Sized Computer*, in "JOBIM - Journées Ouvertes en Biologie, Informatique et Mathématiques", Toulouse, France, July 2013, http://hal.inria.fr/hal-00842027