IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Cachan

Université Rennes 1

Activity Report 2015

Project-Team GENSCALE

Scalable, Optimized and Parallel Algorithms for Genomics

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Computational Biology

Table of contents

Project-Team GENSCALE

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

Keywords:

Computer Science and Digital Science:

3.1.8. - Big data (production, storage, transfer)
3.3. - Data and knowledge analysis
7.1. - Parallel and distributed algorithms

Other Research Topics and Application Domains:

1.1.6. - Genomics
1.1.9. - Bioinformatics
2.2.3. - Cancer

1. Members

Research Scientists

Dominique Lavenier [Team leader, Cnrs, Senior Researcher, HdR]
Claire Lemaitre [Inria, Researcher]
Pierre Peterlongo [Inria, Researcher]
Guillaume Rizk [Associate Researcher, ANR HydroGen]

Faculty Members

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

Engineers

Laurent Bouri [Cnrs, France Génomique]
Sebastien Brillet [Inria, until Mar 2015, Brittany Region]
Erwan Drezen [Inria, until May 2015, ANR GATB]
Patrick Durand [Inria, from Oct 2015]
Anais Gouin [Inria, until Oct 2015, ANR ADA-SPODO]
Fabrice Legeai [Inra]
Sebastien Letort [Cnrs, from Nov 2015]
Ivaylo Petrov [Inria, Brittany Region]
Chloe Riou [Inria, ANR Colib'read]

PhD Students

Mathilde Le Boudic-Jamin [Univ. Rennes I, until Aug 2015]
Gaetan Benoit [Inria, ANR HydroGen]
Cervin Guyomar [Univ. Rennes I, from Oct. 2015]
Antoine Limasset [Univ. Rennes I]
Camille Marchet [Univ. Rennes I, from Oct. 2015]
Francois Moreews [Inra]
Hoang Son Pham [Vietnam gov.]

Post-Doctoral Fellows

Rodrigo Bentes Kato [UFMG, Brazil, until Nov. 2015]
Warley Gramacho Da Silva [Tocantins University, Brazil, from Sep. 2015]

Visiting Scientist

Diep Nguyen [Associate Professor, until Jan 2015]

Administrative Assistant
Marie Le Roic [Univ. Rennes I]

2. Overall Objectives

2.1. Optimization of genomic data processing

The first objective of GenScale is the design of scalable, optimized and parallel algorithms for processing the mass of genomic data provided by today biotechnologies. More specifically, our research activities focus on the optimization of the following treatments:

- Processing of HTS data (High Throughput Sequencing) generated by sequencers of 2nd and 3rd generation. These machines generate billions of short DNA fragments (called reads) requiring treatments such as read compression, read correction, genome assembly (contig generation, scaffolding) and detection of variants (SNPs, breakpoint, inversion, etc.).
- Comparison of large genomic or metagenomic data sets. This fundamental bioinformatics task, due to the steadily increasing of genomic data, is still a bottleneck in many treatments such as taxonomic assignation, functional assignation, genome annotation, etc. Furthermore, the data analyzes of large metagenomic projects don't scale with standard sequence comparison methods. New strategies must be investigated.
- 3D protein structure. Functionalities of proteins are mainly supported by their three dimensional structures. Determining these structures from RMN data or classifying them based on their 3D structures into families require the development of highly optimized algorithms.

Optimization is addressed both in terms of memory space and computation time. Space optimization aims to lower the memory footprint of the algorithms. This is done by the design of innovative data structures. Time optimization aims to provide algorithms with short computation time. Two main ways are followed: combinatorial optimization and multilevel parallelism.

2.2. Active collaboration with life science actors

The second GenScale objective is to create and maintain permanent partnerships with life science research groups. It also aims to be involved in challenging genomic projects in the following areas:

- Health;
- Agronomy and Environment.

GenScale is an interdisciplinary project, which requires strong links with the biology and the genomic scientific community. Hence, it is highly important to keep close relationships with end-users, and being able to have a quick feedback, especially through relevant bioinformatics studies. This is a guarantee for answering right biological questions through right bioinformatics tools.

Collaborations with life science partners go through local, national or international common projects where our tools and methodologies are intensively tested and used. GenScale also welcomes people from INRA, the French research institute in agronomy.

3. Research Program

3.1. Introduction

Based on the overall objectives, the research program of GenScale is structured into four research axes as described below. The first three axes include pure computer science aspects, such as the development of advanced data structures and/or the design of new optimized algorithms; they also include strong partnerships with life science actors to validate the methodologies that are developed. The fourth axis can be seen as a transversal one. It addresses efficient parallel implementations of our methods on standard processors, cluster systems, or accelerators such as GPU.

3.2. Axis 1: HTS data processing

The raw information delivered by NGS (Next Generation Sequencing) technologies represents billions of short DNA fragments. An efficient structuration of this mass of data is the de-Bruijn graph that is used for a large panel of problems dealing with high throughput genomic data processing. The challenge, here, is to represent this graph into memory. An efficient way is to use probabilistic data structures, such as Bloom filters but they generate false positives that introduce noise and may lead to errors. Our approach is to enhance this basic data structure with extra information to provide exact answers, while keeping a minimal memory occupancy [2], [3].

Based on this central data structure, a large panel of HTS algorithms can be designed: read compression, read correction, genome assembly, detection of SNPs (Single Nucleotide Polymorphism) or detection of other variants such as inversion, transposition, etc. [8], [11], [9]. The use of this compact structure guarantees software with very low memory footprint that can be executed on many standard-computing resources.

In the full assembly process, an open problem due to the structure complexity of many genomes is the scaffolding step that consists in reordering contigs along the chromosomes. This treatment can be formulated as a combinatorial optimization problem exploiting the upcoming new sequencing technologies based on long reads.

3.3. Axis 2: Sequence comparison

Comparing genomic sequences (DNA, RNA, protein) is a basic bioinformatics task. Powerful heuristics (such as the seed-extend heuristic used in the well-known BLAST software) have been proposed to limit the computation time. The underlying data structures are based on seed indexes allowing a drastic reduction of the search space. However, due to the increasing flux of genomic sequences, this treatment tends to increase and becomes a critical section, especialy in metagenomic projects where hundred of millions of reads must be compared to large genomic banks for taxonomic of functional assignation.

Our research follows mainly two directions. The first one revisits the seed-extend heuristic in the context of the bank-to-bank comparison problem. It requires new data structures to better classify the genomic information, and new algorithmic methods to navigate through this mass of data [7], [10]. The second one addresses metagenomic challenges that have to extract relevant knowledge from Tera bytes of data. In that case, the notion of sequence similarity itself is redefined in order to work on objects that are much simpler than the standard alignment score, and that are better suited for large-scale computation. Raw information (reads) is first reduced to k-mers from which high speed and parallel algorithms compute approximate similarities based on a well defined statistical model [4].

3.4. Axis 3: Protein 3D structure

The three-dimensional (3D) structure of proteins tends to be evolutionarily better preserved during evolution than its sequence. Finding structural similarities between proteins gives deep insights into whether these proteins share a common function or whether they are evolutionarily related. Structural similarity between two proteins is usually defined by two functions – a one to- one mapping (also called alignment) between two subchains of their 3D representations and a specific scoring function that assesses the alignment quality. The structural alignment problem is to find the mapping that is optimal with respect to the scoring function. Protein structures can be represented as graphs, and the problem reduces to various combinatorial optimization problems that can be formulated in this framework: for example finding the maximum weighted path [1] or finding the maximum cardinality clique/pseudo-clique [5] [18].

In most cases, however, suitable conformations for a given protein are unknown. To support this statement, we point out that the number of deposited protein conformations on the Protein Data Bank (PDB [1]) recently reached the threshold of 110,000 entries, while the UniProtKB/TrEMBL [2] database contains more than 50

[1] http://www.rcsb.org/
[2] http://www.ebi.ac.uk/uniprot/TrEMBLstats

million sequence entries, all of them potentially capable for coding for a new protein. In this context, distance geometry provides powerful methods and algorithms for the identification of protein conformations from Nuclear Magnetic Resonance (NMR) data, which basically consist of a distance list concerning atom pairs of the protein[6]. We are working on the discretization of the distance geometry, so that its search space becomes discrete (and finite!), for making it possible to perform an exhaustive exploration of the solution set.

3.5. Axis 4: Parallelism

Together with the design of new data structures and new algorithms, our research program aims to propose efficient hardware implementation. Even if not explicitly mentioned in the three previous axes, we have constantly in mind to exploit the parallelism of current processors. Practically, and depending of the nature of the computation to perform, three levels of parallelism are addressed: the use of vector instructions of today processors, the multithreading offered by multi-core systems, and the cluster (or cloud) infrastructures.

Consequent bioinformatics treatments, from the processing of raw HTS data to high-level analysis, are generally performed within a workflow environment and executed on cluster systems. Automating the parallelization of such treatments directly from a graphical capture of the workflow is a necessity for end-users that are generally not expert in parallelism. The challenge here is to hide, as much as possible, the different transformations to go from a high level workflow description to an efficient parallel execution that exploits both task-level and data-level parallelism[25].

Another research activity of this axe is the design of parallel algorithms targeting hardware accelerators, especially GPU boards (Graphical Processing Unit). These devices now offer a high-level programming environment to access the hundred of processors available on a single chip [20]. A few bioinformatics treatments, such as ones that exhibit good computational regularity, can highly benefit from the computing power of this technology.

4. Application Domains

4.1. Introduction

Today, sequencing data are intensively used in many life science projects. The methodologies developed by the GenScale group are generic approaches that can be applied to a large panel of domains such as health, agronomy or environment areas. The next sections briefly describe examples of our activity in these different domains.

4.2. Health

Cancer diagnostic: from a pool of known genes, the aim is to detect potential mutations that perturb the activity of these genes. Pointing out the right gene help in prescribing the right drug. The bioinformatics analysis is based on the detection of SNPs (Single Nucleotide Polymorphism) from a set of target genes.

Microbiology: Streptococcus bacteria are considered as major pathogens for humans and lead to many infections. The cause of their pathogenicity can be studied from their genomic structure by comparing different strains. Text of the genomes must first be constructed (assembly process) before to be analyzed (comparative genomic).

HLA genotyping: The human leukocyte antigen (HLA) system drives the regulation of the Human immune system. The HLA genes reside on chromosome 6 and have a large number of alleles. Genotyping this group of genes can be done by a deep sequencing of the HLA region, and by comparing reads with a HLA databank (intensive sequence comparison).

4.3. Agronomy and Environment

Improving plant breeding: such projects aims at 1) identifying favorable alleles at loci contributing to phenotypic variation, 2) characterizing N-traits at the functional level and 3) providing robust multi-locus SNP-based predictors of the breeding value of agronomical traits under polygenic control. Underlying bioinformatics processing is the detection of informative zones (QTL) on the plant genomes.

Insect study: Insects represent major crop pests, justifying the need for control strategies to limit population outbreaks and the dissemination of plant viruses they frequently transmit. Several issues are investigated through the analysis and comparison of their genomes: understanding their phenotypic plasticity such as their reproduction mode changes, identifying the genomic sources of adaptation to their host plant and of ecological speciation, and understanding the relationships with their bacterial symbiotic communities.

Ocean biodiversity: The metagenomic analysis of seawater samples provides an original way to study the ecosystems of the oceans. Through the biodiversity analysis of different ocean spots, many biological questions can be addressed, such as the plankton biodiversity and their role, for example, in the CO2 sequestration.

5. Highlights of the Year

5.1. Highlights of the Year

Special Issue

Publication of a special issue on Discrete Applied Mathematics. Guest Editors: A. Mucherino, R. de Freitas, C. Lavor [35]

Awards

For the third time in the last three editions of JOBIM (National workshop on Biology, Informatics and Mathematics), PhD students of the GenScale team won the best poster award:

- JOBIM 2015: Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets [39] (https://hal.inria.fr/hal-01180603)
- JOBIM 2013 : MINIA on a Raspberry Pi, Assembling a 100 Mbp Genome on a Credit Card Sized Computer (https://hal.inria.fr/hal-00842027)
- JOBIM 2012 : Compareads: comparing huge metagenomic experiments (https://hal.inria.fr/hal-00760332)

In 2014, due to the ECCB conference in Strasbourg, France, there was no specific JOBIM event.

6. New Software and Platforms

6.1. HTS data processing

6.1.1. GATB: Genome Assembly & Analysis Tool Box

The GATB software toolbox aims to lighten the design of NGS algorithms. It offers a panel of high-level optimized building blocks to speed-up the development of NGS tools related to genome assembly and/or genome analysis. The underlying data structure is the de Bruijn graph, and the general parallelism model is multithreading. The GATB library targets standard computing resources such as current multicore processors (laptop computer, small server) with a few GB of memory. From high-level API, NGS programming designers can rapidly elaborate their own software based on domain state-of-the-art algorithms and data structures. The GATB library is written in C++.

Contact: Dominique Lavenier

URL: https://gatb.inria.fr/

6.1.2. LEON: Genomic Data Compression

Leon is a lossless compression software that achieves compression of DNA sequences of high throughput sequencing data, without the need of a reference genome. Techniques are derived from assembly principles that better exploit NGS data redundancy. A reference is built de novo from the set of reads as a probabilistic de-Bruijn graph stored in a Bloom filter. Each read is encoded as a path in this graph, storing only an anchoring kmer and a list of bifurcations indicating which path to follow in the graph. This new method will allow to have compressed read files containing its underlying de-Bruijn Graph, thus directly re-usable by many tools relying on this structure. Leon achieved the encoding of a *C. elegans* reads set with 0.7 bits per base, outperforming state of the art reference-free methods.

Contact: Claire Lemaitre

URL: https://gatb.inria.fr/software/leon/

6.1.3. BLOOCOO: Genomic Data Correction

Bloocoo is a k-mer spectrum-based read error corrector, designed to correct large datasets with a very low memory footprint. It uses the disk streaming k-mer counting algorithm included in the GATB library, and inserts solid k-mers in a bloom-filter. The correction procedure is similar to the Musket multistage approach. Bloocoo yields similar results while requiring far less memory: as an example, it can correct whole human genome re-sequencing reads at 70 x coverage with less than 4GB of memory.

Contact: Claire Lemaitre

URL: https://gatb.inria.fr/bloocoo-read-corrector/

6.1.4. DiscoSnp++: DISCOvering Single Nucleotide Polymorphism

DiscoSnp++ is designed for discovering Single Nucleotide Polymorphism (SNP) and insertions/deletions (indels) from raw set(s) of reads obtained with Next Generation Sequencers (NGS). The number of input read sets is not constrained, it can be one, two, or more. No other data as reference genome or annotations are needed. The software is composed of three modules: (1) kissnp2, that detects SNPs and indels from read sets; (2) kissreads2, that enhances the kissnp2 results by providing for each variant a read coverage mean and a (phred) quality; (3) VCF_creator, that provides a file in the Variant Calling Format (VCF). A VCF file using or not a reference genome is also created.

Contact: Pierre Peterlongo

URL: http://colibread.inria.fr/software/discosnp/

6.1.5. MindTheGap: Detection of insertion

MindTheGap is a software that performs detection and assembly of DNA insertion variants in NGS read datasets with respect to a reference genome. It takes as input a set of reads and a reference genome. It outputs two sets of FASTA sequences: one is the set of breakpoints of detected insertion sites, the other is the set of assembled insertions for each breakpoint. For each breakpoint, MindTheGap either returns a single insertion sequence (when there is no assembly ambiguity), or a set of candidate insertion sequences (due to ambiguities) or nothing at all (when the insertion is too complex to be assembled). MindTheGap performs de novo assembly using the de Bruijn Graph implementation of GATB. Hence, the computational resources required to run MindTheGap are significantly lower than that of other assemblers.

Contact: Claire Lemaitre

URL: http://mindthegap.genouest.org/

6.1.6. TakeABreak: Detection of inversion breakpoints

TakeABreak is a tool that can detect inversion breakpoints directly from raw NGS reads, without the need of any reference genome and without de novo assembling the genomes. Its implementation is based on the Genome Assembly Tool Box (GATB) library, and has a very limited memory impact allowing its usage on common desktop computers and acceptable runtime (Illumina reads simulated at 80x coverage from human chromosome 22 can be treated in less than two hours, with less than 1GB of memory).

Contact: Claire Lemaitre

URL: http://colibread.inria.fr/software/takeabreak/

6.2. Sequence comparison

6.2.1. PLAST: Parallel Local Alignment Search Tool

PLAST is a fast, accurate and NGS scalable bank-to-bank sequence similarity search tool providing significant accelerations of seeds-based heuristic comparison methods, such as the Blast suite. PLAST is fully designed to compare query and subject comprised of large sets of DNA, RNA and protein sequences. It is significantly faster than BLAST, while providing comparable sensitivity. PLAST contains a fully integrated data-filtering engine capable of selecting relevant hits with user-defined criteria (E-Value, identity, coverage, alignment length, etc.).

Contact: Dominique Lavenier

URL: https://plast.inria.fr

6.2.2. SIMKA: Comparison of metagenomic datasets

Simka rapidly compares a large number of metagenomics datasets using efficient kmer-based method. Datasets may contains hundreds of millions of NGS sequences. Kmers of each datasets are rapidly counted and identified to estimate the pairwise similarities between datasets. The output of Simka can be used for clustering purpose or for checking correlation between metadata.

Contact: Gaëtan Benoit

URL: https://gatb.inria.fr/software/simka/

6.2.3. BGREAT: read mapper on de-Bruijn graph

BGREAT maps reads on a de-Bruijn Graph, usually used for genome assembly. Mapping reads on graphs offers the possibility to conserve all the pieces of information lost during the assembly process and to avoid multi-mapping problems due to genomic repeats. BGREAT rewrites a read sequence as a succession of unitigs sequences. It can map millions of reads per CPU hour on a de-Bruijn graph built from a large set of human genomic reads.

Contact: Antoine Limasset

URL: http://github.com/Malfoy/BGREAT/

7. New Results

7.1. HTS data processing

7.1.1. Genome Analysis Tool Box Optimization

Participants: C. Deltel, P. Durand, E. Drezen, D. Lavenier, C. Lemaitre, P. Peterlongo, G. Rizk

Among the GATB library, the kmer-counting procedure is one of the most useful building block to speed-up development of new NGS tools. It is the first step of many NGS tools developed with GATB : Leon, Bloocoo, MindTheGap, DiscoSnp, Simka, TakeAbreak. This procedure has been optimized to be less limited by disks I/O. It relies on the use of kmer minimizers that help quickly partition the whole set of kmers into compact subsets. The kmer-counting procedure has also been re-worked to be more versatile, it is now able to count separately many input files and allows easy parametrization of the output, from simple kmer-count to the creation of custom user-defined kmer measures. At the core of the GATB library is also the manipulation and traversal of the de Bruijn Graph. The implementation has been optimized, leading to graph traversal twice fast as before. We introduced a new type of bloom filters, that are specially optimized for the manipulation of kmers. In these bloom filters neighboring kmers in the graph are close together in the bloom filter bit array, leading to better data locality, less cache misses and better overall performance [38].

7.1.2. NGS Data Compression

Participants: G. Benoit, E. Drezen, D. Lavenier, C. Lemaitre, G. Rizk

A novel reference-free method to compress data issued from high throughput sequencing technologies has been developed. Our approach, implemented in the LEON software, employs techniques derived from assembly principles. The method is based on a reference probabilistic de-Bruijn Graph, built de novo from the set of reads and stored in a Bloom filter. Each read is encoded as a path in this graph, by memorizing an anchoring kmer and a list of bifurcations. The same probabilistic de-Bruijn Graph is used to perform a lossy transformation of the quality scores allowing higher compression rates to be obtained without loosing pertinent information for downstream analyses. Leon was run on various real sequencing datasets (whole genome, exome, RNA-seq or metagenomics). In all cases, LEON showed higher overall compression ratios than state-of-the-art compression software. On a *C. elegans* whole genome sequencing dataset, LEON divided the original file size by more than 20 [16].

7.1.3. Multistep global optimization approach for the scaffolding problem

Participants: R. Andonov, D. Lavenier, I. Petrov

Our overall goal here is to address the computational hardness of the scaffolding problem by designing faster algorithms for global optimization that combine the branch-and-bound method which is able to find the global optimum but is usually slow for accuracy, with the use of massive parallelism and exploiting the special properties of the data–for scalability. A new two step scaffolding modeling strategy is in development. It tries to break the problem complexity by first solving a graph containing only large unitigs building something that can be compared to a trustworthy genomic frame. In our preliminary works [40] we developed integer programming optimization models that have been successfully applied on synthetic data generated from small chloroplast genomes. For computation we uses the Gurobi optimization solver.

7.1.4. Mapping reads on graph

Participants: A. Limasset, C. Lemaitre, P. Peterlongo

Next Generation Sequencing (NGS) has dramatically enhanced our ability to sequence genomes, but not to assemble them. In practice, many published genome sequences remain in the state of a large set of contigs. Although many subsequent analyses can be performed, one may ask whether mapping reads on the contigs is as informative as mapping them on the paths of the assembly graph. We proposed a formal definition of mapping on a de Bruijn graph, analysed the problem complexity which turned out to be NP-complete, and provided a practical solution. We proposed a pipeline called GGMAP (Greedy Graph MAPping). Its novelty is a procedure to map reads on branching paths of the graph, for which we designed a heuristic algorithm called BGREAT (de Bruijn Graph REAd mapping Tool). For the sake of efficiency, BGREAT rewrites a read sequence as a succession of unitigs sequences. GGMAP can map millions of reads per CPU hour on a de Bruijn graph built from a large set of human genomic reads. Surprisingly, results show that up to 22% more reads can be mapped on the graph but not on the contig set. Although mapping reads on a de Bruijn graph is a complex task, our proposal offers a practical solution combining efficiency with an improved mapping capacity compared to assembly-based mapping even for complex eukaryotic data [43].

7.1.5. Improving discoSnp features

Participants: C. Riou, C. Lemaitre, P. Peterlongo

NGS data enable to detect polymorphisms such as SNPs and indels. Their detection in NGS data is now a routine task. The main methods for their prediction usually need a reference genome. However, non-model organisms and highly divergent genomes such as in cancer studies are more and more investigated. The discoSnp tool has been successfully applied to predict isolated SNPs from raw read set(s) without the need of a reference genome. We improved discoSnp which became discoSnp++ [44]. DiscoSnp++ benefits from a new software design that reduces time and memory consumption, and from a new algorithmic design that detects all kinds of SNP and small indels, adds genotype information and outputs a VCF (Variant Calling Format) file. Moreover, when a reference genome may be used, discoSnp++ predictions are automatically mapped to this reference and the VCF file shows up location information of each prediction. This step also provides a way to filter out false predictions due to genomic repeats. Using discoSnp++ even when a reference is available has multiple advantages: it is several order of magnitude faster and uses much less memory. We are currently working in showing that it also provides better predictions than methods based on read mapping.

7.1.6. HLA genotyping

Participant: D. Lavenier

The human leukocyte antigen (HLA) system drives the regulation of the Human immune system. Genotyping the HLA genes involved in the immune system consists first in a deep sequencing of the HLA region. Next, a NGS analysis is performed to detects SNP variations from which correct haplotypes are computed. We have developed a fast method that outperforms standard approaches which, generally, require exhaustive database searches. Instead, the method extracts a few significant k-mers from all the haplotypes referenced in the HLA database. Each haplotype is then characterized by a small set of informative k-mers. By comparing these k-mer sets with the HLA sequencing data of a specific person, we can rapidly determine its HLA genotype.

7.1.7. Identification of long non-coding RNAs in insects genomes

Participant: F. Legeai

The development of high throughput sequencing technologies (HTS) has allowed researchers to better assess the complexity and diversity of the transcriptome. Among the many classes of non-coding RNAs (ncRNAs) that were identified during the last decade, long non-coding RNAs (lncRNAs) represent a diverse and numerous repertoire of important ncRNAs, reinforcing the view that they are of central importance to the cell machinery in all branches of life. Although lncRNAs have been involved in essential biological processes such as imprinting, gene regulation or dosage compensation especially in mammals, the repertoire of lncRNAs is poorly characterized for many non-model organisms [23]. In collaboration with the Institut de Génétique et de Développement de Rennes (IGDR) we participate in the development of a software for extracting long non coding RNA from high throughput data (https://github.com/tderrien/FEELnc).

7.1.8. Data-mining applied to GWAS

Participants D. Lavenier, Pham Hoang Son

Discriminative pattern mining methods are powerful techniques for discovering variant combinations related to diseases. The aim is to find a set of patterns that occur with disproportion frequency in case-control data sets, and a real challenge is to select a complete set of variant combinations that are biologically significant. There are various measurement methods for evaluating the discriminative power of individual combination in two-class data sets. Our research activity on this topic attempts to compare the statistical discriminative power measurements in genetic case-control data sets in order to evaluate the effectiveness of detecting variants associated with diseases.

7.2. Sequence comparison

7.2.1. Amplicon alignment

Participants: S. Brillet, C. Deltel, P. Durand, D. Lavenier, I. Petrov

Many metagenomics projects identify species by the studying16S-RNA sequences. This is mainly done by comparing the amplicons with 16S-RNA bacterial banks (amplicons are short fragments sequenced from very specific genome areas). As these sequences share a lot of similarities, immediate blast-like heuristics achieve poor performances. To speed up the process, we first select informative k-mers, from both the amplicon dataset and in the RNA16 bank (informative k-mers are defined as under represented k-mers). An index is built from this reduced set of k-mers and a "seed-and-extend" procedure is run. This strategy avoids many non-useful computation and accelerate the overall computation by two orders of magnitude. This new approach is currently implemented in the PLAST software (Regional KoriPlast2 project).

7.2.2. Metagenomics datasets comparison

Participants: G. Benoit, D. Lavenier, C. Lemaitre, P. Peterlongo, G. Rizk

We developp a new method, called Simka, to compare simultaneously numerous large metagenomics datasets. The method computes pairwise distances based on the amount of shared k-mers between datasets. The method scales to a large number of datasets thanks to an efficient kmer-counting step that processes all datasets simultaneoulsy. Additionnally, several distance definitions were implemented and compared, including some originating from the ecological domain. The method is currently applied to the TARA oceans project (more than 500 datasets) which aims at comparing worldwide sea water samples (ANR HydrGen project) [39].

7.3. Protein 3D structure

7.3.1. Discovering protein conformations by distance geometry

Participant: A. Mucherino

The distance geometry asks whether a simple weighted undirected graph G can be embedded in a Euclidean space having a predefined dimension $K > 0$, so that distances between pairs of embedded vertices are the same as the weights on graph edges. One of the most important applications of the distance geometry can be found in biology, where experimental techniques are able to find estimates of certain distances between atom pairs in molecules. Even if the scientific community is used to employ standardized techniques for the solution of this problem, which are essentially based on heuristic searches, we have recently shown that our combinatorial approach to this problem can be in fact employed for solving biological instances of the distance geometry [17]. This work is in collaboration with international people and researchers from the Pasteur Institut in Paris.

7.3.2. Discretization orders for distance geometry

Participant: A. Mucherino

The concept of discretization order is fundamental for the discretization of the distance geometry, i.e. for reducing the search space of a given distance geometry instance to a discrete (and finite) space. A discretization order is an order on the vertices of the graph G representing an instance of the distance geometry that is able to satisfy the discretization assumptions. Recent research was focused on the problem of finding, for a given distance geometry instance, a suitable discretization order that allows for its discretization [32]. The problem is tackled from a purely theoretical point of view in [33], while a special order for protein backbones was identified in [27] by creating a path on a "pseudo" de Bruijn graph. In [36], additional requirements are included during the search for a vertex order, in order to identify discretization orders that are also "optimal". In this work, we used Answer Set Programming (ASP) for identifying optimal partial orders that ensure the discretization of distance geometry instances related to proteins. This work is in collaboration with the Dyliss team, as well as international people.

7.3.3. Structure Similarity Detection

Participants: M. Le Boudic-jamin, R. Andonov

The most commonly used among the various measures of alignment similarity are the internal distances root mean squared deviation (RMSDd) and the coordinate root mean squared deviation (RMSDc) . In the paper [18] we introduce a novel approach to find similarities between protein structures. Our algorithm is both internal-distances based and Euclidean-coordinates based (i.e., it uses a rigid transformation to optimally superimpose the two structures). Resulting alignments are guaranteed to score well for both RMSDd and RMSDc , while remaining polynomial. We also replace the goal of finding the largest clique by the one of returning several very dense "near-clique" subgraphs. This choice is strongly justified by the observation that distinct solutions to the structural alignment problem that are close to the optimum are all equally viable from the biological perspective, and hence are all equally interesting from the computation standpoint. Our tool is suitable for detecting similar domains when comparing multi-domain proteins, as well to detect structural repetitions within a single protein and between related proteins [12].

7.3.4. Automatic Classification of Protein Structure

Participants: M. Le Boudic-jamin, R. Andonov

In this paper [15] we propose a new distance measure for comparing two protein structures based on their contact map representations . We show that our novel measure, which we refer to as the maximum contact map overlap (max-CMO) metric, satisfies all properties of a metric on the space of protein representations. Having a metric in that space allows one to avoid pairwise comparisons on the entire database and, thus, to significantly accelerate exploring the protein space compared to no-metric spaces. We show on a gold standard superfamily classification benchmark set of 6759 proteins that our exact k-nearest neighbor (k-NN) scheme classifies up to 224 out of 236 queries correctly and on a larger, extended version of the benchmark with 60 850 additional structures, up to 1361 out of 1369 queries. Our k-NN classification thus provides a promising approach for the automatic classification of protein structures based on flexible contact map overlap alignments.

7.3.5. Detection of structure repeats in proteins

Participant: M. Le Boudic-jamin, R. Andonov

Almost 25% of proteins contain internal repeats, these repeats may have a major role in the protein function. Furthermore some proteins actually are the same substructure repeated many times, these proteins are solenoids. However, very few protein repeats detection programs exist today. In the paper [29] we present a simple and efficient tool for discovering protein repeats. Our tool is based on protein fragment comparison and clique detection. We show that our tool is able to detect different levels of repetitions and to successfully identify protein tiles.

7.4. Parallelism

7.4.1. Processor in Memory

Participants: C. Deltel, D. Lavenier

The concept of PIM (Processor In Memory) aims to dispatch the computer power near the data. Together with the UPMEM company, which is currently developing a DRAM enhanced with computing units, we investigate the parallelization of several bioinformatics algorithms for this new types of memory. The first results show that blast-like algorithms or mapping algorithms can highly benefit of such memory. But the core algorithms must be revisited in order to better suite the PIM architecture.

7.4.2. Alignment search tools on cloud

Participants: S. Brillet, D. Lavenier, I. Petrov

PLAST is an alternative version of Blast to target intensive sequence comparison (bank-to-bank comparison). The multicore version offers a speed from 5 to 10 compared to Blast. In 2015, we deploy PLAST in the IFB cloud infra-structure (French Bioinformatics Institute) and demonstrate that an Hadoop implementation provides a very good scalability [34].

7.4.3. Bioinformatics Workflow

Participants: D. Lavenier, F. Moorews

Bioinformatics workflows play an important role in the development of new methodologies for analyzing sequencing data. Optimizing this activity brings the questions of how workflow can be efficiently captured and how technical tasks integration can be simplified. Thus, we define an expressive graphic worfklow language, adapted to the quick capture of workflows. This graphical input is then interpreted by a workflow engine based on a new model of computation with high performances obtained by the use of multiple levels of parallelism. A Model-Driven design approach is associated to facilitate the data parallelism generation and the production of suitable implementations for different execution contexts. In the case of the cloud model Container as a Service (CaaS), a workflow specification intrinsically re-executable and readily disseminatable has been developed. The adoption of this kind of model could lead to an acceleration of exchanges and a better availability of data analysis workflows [25] [31] [13].

7.4.4. Graph processing

Participants: D. Lavenier, R. Andonov

In the paper [20] we present a new approach for solving the all-pairs shortest-path (APSP) problem for planar graphs that exploits the massive on-chip parallelism available in today's Graphics Processing Units (GPUs). We describe two new algorithms based on our approach. Both algorithms use Floyd-Warshall method, have near optimal complexity in terms of the total number of operations, while their matrix-based structure is regular enough to allow for efficient parallel implementation on the GPUs. By applying a divide-and-conquer approach, we are able to make use of multi-node GPU clusters, resulting in more than an order of magnitude speedup over fastest known Dijkstra-based GPU implementation and a two-fold speedup over a parallel Dijkstra-based CPU implementation.

7.4.5. Analytical models and Optimization for GPUs

Participants: R. Andonov

In [28] we develop a methodology for modeling the energy efficiency of tiled nested-loop codes running on a graphics processing unit (GPU) and use it for energy efficiency optimization. We use the polyhedral model, and we assume that a highly optimized and parametrized version of a tiled nested – loop code, either written by an expert programmer or automatically produced by a polyhedral compilation tool – is given to us as an input. We then model the energy consumption as an analytical function of a set of parameters characterizing the software and the GPU hardware. Our approach develops analytical models based on (i) machine and architecture parameters, (ii) program size parameters as found in the polyhedral model and (iii) tiling parameters, such as those that are chosen by auto-or manual tuners. Our model therefore allows efficient optimization of the energy efficiency with respect to a set of parameters of interest.

7.5. Applications

7.5.1. CAMI: Critical Assessment of Metagenomic Interpretation

Participants: C. Deltel, D. Lavenier, C. Lemaitre, P. Peterlongo, G. Rizk

The interpretation of metagenomes relies on sophisticated computational approaches such as short read assembly, binning and taxonomic classification. All subsequent analyses can only be as meaningful as the outcome of these initial data processing methods [3]. The CAMI initiative aims to evaluate these methods independently, comprehensively and without bias. The goal is to supply users with exhaustive quantitative data about the performance of methods in many relevant scenarios. In 2015, we participate to CAMI within the "assembly" category using the Minia assembly pipeline. Results are provided here: https://data.cami-challenge.org/. For the medium challenge datasets, our assemblies are refered under the identifiers *goofy-wilson* and *fervent-blackwell*.

[3] http://www.cami-challenge.org/

7.5.2. Assembly and Annotation of Arthropods Genomes

Participants: A. Gouin, F. Legeai, C. Lemaitre

Within a large international network of biologists, GenScale has contributed to various projects for identifying important components such as protein coding or non coding genes involved in the adaptation of major agricultural pests to their environment. We provided the assembly and the annotation of 4 new aphids, 3 parasitic wasps, and improved the assembly of 2 variants of fall army worm by removing unwanted sequences due to heterozygosity [41], [42]. Following specific agreement or policy, these new genomes and annotations are available for a restricted consortium or a large community through the BioInformatics platform for Agro-ecosystems Arthropods (http://bipaa.genouest.org/is). These results, and further analyses led to a better understanding of the biology, evolution and life history traits of *Spodoptera frugiperda* [19], the identiifcation and characterization of new genome of pea aphid symbionts [22] and the identification of differentially expressed genes in the sensory system of Sesamia nonagrioides [21].

7.5.3. Study of the rapeseed genome structure

Participants: D. Lavenier, C. Lemaitre, S. Letort, P. Peterlongo

In collaboration with IGEPP (Institut de Génétique, Environnement et Protection des Plantes), INRA, and through two national projects, PIA Rapsodyn and France-Génomique Polysuccess, we are involved in the genome analysis of several rapeseed varieties. The Rapsodyn project has the ambition to insure long-term competitiveness of the rapeseed production through improvement of the oil yield and reduction of nitrogen inputs during the crop cycle. Rapeseed varieties must thus be selected from genotypes that favor low nitrogen input. DiscoSNP++ is here used to locate new variants among the large panel of rapeseed varieties which have been sequenced during the project. The PolySuccess project aims to answer the following question: how a polyploid, such as the oilseed rape plant, becomes a new species? Oilseed rape (Brassica napus) beeing a natural hybrid between B.rapa and B.oleracea, different genomes of these three species have been sequenced to study their structures. The Minia assembly pipeline provides a fast way to generate contigs that are used for studying gene specificities.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Empowered memory

Participants: Charles Deltel, Dominique Lavenier.

The UPMEM company is currently developing new memory devices with embedded computing power. GenScale investigates how bioinformatics algorithms can benefit from these new types of memory (see section New Results).

8.2. Bilateral Grants with Industry

8.2.1. Korilog: I-Lab KoriScale

Participants: Sébastien Brillet, Erwan Drezen, Dominique Lavenier, Ivaylo Petrov.

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 enabled 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. In June 2015, Korilog stopped its activity.

8.2.2. Rapsodyn project

Participants: Dominique Lavenier, Claire Lemaitre, Sebastien Letort, Pierre Peterlongo.

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 optimization of the rapeseed oil content and yield under low nitrogen input. GenScale is involved in the bioinformatics work package, in collaboration with Biogemma's bioinformatics team, to elaborate advanced tools dedicated to polymorphism.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. Bioinformatics computing center of Roscoff

Participants: Sébastien Brillet, Erwan Drezen, Patrick Durand, Dominique Lavenier, Ivaylo Petrov.

Through the collaborative project KORIBLAST2 funded by Région Bretagne (June 2014-December 2015) and within the KoriScale lab, we worked: (1) to improve the KLAST software with new alignment methods developed by GenScale; (2) to extend the capabilities of KLAST toward metagenomic processing; (3) to develop a cloud version targeting huge sequence comparison processing.

9.1.2. Etablissement Français du sang (EFS)

Participant: Dominique Lavenier.

An active collaboration with EFS started in 2015 to speed up individual HLA genotyping. A first prototype has been designed (see section New Results) and should be intensively tested in 2016 on many patient data.

9.1.3. Rennes Hospital, Hematology service, Genetic service

Participants: Patrick Durand, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk.

The collaboration with the Hematology service and with the Genetic service of the Rennes hospital aims to set up advanced bioinformatics pipelines for cancer diagnosis.

9.1.4. Partnership with INRA in Rennes

Participants: Cervin Guyomar, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, Sébastien Letort, Pierre Peterlongo, François Moreews.

The GenScale team has 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 two INRA engineers (F. Legeai, F. Moreews) and one PhD student (C. Guyomar).

9.2. National Initiatives

9.2.1. ANR

9.2.1.1. Project FATINTEGER

Participants: Dominique Lavenier, François Moreews.

Coordinateur: F. Gondret
Duration: 36 months (Mar. 2012 - feb. 2015)
Partners: PEGASE Inra Rennes, CNRS IRISA Rennes, AgroCampus Ouest LMA-IRMAR Rennes

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. GenScale is involved in the design of the workflow for processing the genomic data.

9.2.1.2. Project ADA-SPODO: Genetic variation of Spodoptera Frugiperda

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

Coordinator: E. D'Alençon (Inra, Montpellier)
Duration: 45 months (Oct. 2012 – May 2016)
Partners: DGIMI Inra Montpellier, CBGP Inra Montpellier, URGI Inra Versailles, Genscale Inria/IRISA Rennes.

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 postgenomic and regulatory data.

9.2.1.3. Project COLIB'READ: Advanced algorithms for NGS data

Participants: Pierre Peterlongo, Antoine Limasset, Camille Marchet, Claire Lemaitre, Dominique Lavenier, Fabrice Legeai, Guillaume Rizk, Chloé Riou.

Coordinator: P. Peterlongo (Inria, GenScale, Rennes)
Duration: 45 months (Mar. 2013 – Dec. 2016)
Partners: LIRMM Montpellier, Erable Inria Lyon, Genscale Inria/IRISA Rennes.

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. The goal is to avoid the assembly step that often leads to a significant loss of information, or generates chimerical results due to complex heuristics. Instead, the strategy proposes a set of innovative approaches that bypass the assembly phase, and that does not require the availability of a reference genome. https://colibread.inria.fr/

9.2.1.4. Project GATB: Genome Analysis Tool Box

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

Coordinator: D. Lavenier (Inria/Irisa, GenScale, Rennes)
Duration: 24 months (Feb. 2013 – Jan. 2015)
Partners: GenScale Inria/IRISA, Rennes – DTI Inria, Rennes.

This project aims to develop algorithms and tools for genome analysis based on a compact data structure having a very low memory footprint allowing end-users to process huge volume of genomic data on a simple desktop computer. The GATB is structured around a C++ library from which many efficient NGS tools can be developed. GATB has been published and is used outside Genscale (LIRMM, Inria Erable team). http://gatb.inria.fr

9.2.1.5. Project HydroGen: Metagenomic applied to ocean life study

Participants: Dominique Lavenier, Pierre Peterlongo, Claire Lemaitre, Guillaume Rizk, Gaëtan Benoit.

Coordinator: D. Lavenier (Inria/Irisa, GenScale, Rennes)
Duration: 42 months (Nov. 2014 – Apr. 2018)
Partners: CEA (GenosScope, Evry), INRA (AgroParisTech, Paris – MIG, Jouy-en-Jossas).

The HydroGen project aims to design new statistical and computational tools to measure and analyze biodiversity through comparative metagenomic approaches. The support application is the study of ocean biodiversity based on the analysis of seawater samples available from the Tara Oceans expedition.

9.2.1.6. Project SpeCrep: speciation processes in butterflies

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

Coordinator: M. Elias (Museum National d'Histoire Naturelle, Institut de Systematique et d'Evolution de la Biodiversite, Paris)
Duration: 48 months (Jan. 2015 – Dec. 2018)
Partners: MNHN (Paris), INRA (Versailles-Grignon), Genscale Inria/IRISA Rennes.

The SpeCrep project aims at better understanding the speciation processes, in particular by comparing natural replicates from several butterfly species in a suture zone system. GenScale's task is to develop new efficient methods for the assembly of reference genomes and the evaluation of the genetic diversity in several butterflies populations.

9.2.2. PIA: Programme Investissement d'Avenir

9.2.2.1. RAPSODYN: Optimization of the rapeseed oil content under low nitrogen

Participants: Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

Coordinator: N. Nessi (Inra, IGEPP, Rennes)
The objective of the Rapsodyn project is the optimization of the rapeseed oil content and yield under low nitrogen input. GenScale is involved in the bioinformatics work package to elaborate advanced tools dedicated to polymorphism and application to the rapeseed plant.

9.2.2.2. France Génomique: Bio-informatics and Genomic Analysis

Participants: Laurent Bouri, Dominique Lavenier.

Coordinator: J. Weissenbach (Genoscope, Evry)
France Génomique gathers resources from the main French platforms in genomic and bio-informatics. It offers to the scientific community an access to these resources, a high level of expertise and the possibilities to participate in ambitious national and international projects. The GenScale team is involved in the work package "assembly" to provide expertise and to design new assembly tools for the 3rd generation sequencing.

9.3. International Initiatives

9.3.1. Brazil

- IMECC, UNICAMP, Campinas [A. Mucherino]
- Federal University of Florianópolis, Santa Catarina: Distance geometry, optimal vertex orders [A. Mucherino]

9.3.2. Chile

- university of Utalca, genomes of aphid parasitoids [F. Legeai]

9.3.3. USA

- Los Alamos National Laboratory (LANL), Los Alamos: Graph algorithms, Parallelism, GPU [R. Andonov, D. Lavenier]
- University of Miami, member of the international Aphid genome consortium [F. Legeai]
- University of Arizona, genomes of aphid parasitoids [F. Legeai]
- University of Ohio, genomics of the soybean aphid [F. Legeai]

9.3.4. China

- SouthWest university, member of the international Spodoptera litura genome project [F. Legeai]

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. General chair, scientific chair

- Workshop on Computational Optimization, Lodz, Poland [A. Mucherino]

10.1.2. Scientific events selection

10.1.2.1. Member of the conference program committees

- BIBM 2015: IEEE International Conference on Bioinformatics and Biomedicine [D. Lavenier]
- HPiC 2015: EEE International Conference on High Performance Computing [D. Lavenier]
- PBC 2015: Workshop on Parallel Computational Biology [D. Lavenier]
- RECOMB-SEQ 2015: RECOMB Satellite Workshop on Massively Parallel Sequencing [D. Lavenier]
- WABI 2015: Workshop on Algorithms for Bioinformatics [D. Lavenier]
- BIOKDD 2015: Biological Knowledge Discovery and Data Mining [D. Lavenier]

10.1.2.2. Reviewer

- WABI 2015 [C. Lemaitre]

10.1.3. Journal

10.1.3.1. Member of the editorial boards

- Discrete Applied Mathematics [A. Mucherino]

10.1.3.2. Reviewer - Reviewing activities

- Advances in Bioinformatics [D. Lavenier]
- Algorithms for Molecular Biology [D. Lavenier]
- Bioinformatics [C. Lemaitre, D. Lavenier]
- BMC Bioinformatics [D. Lavenier]
- BMC Genomics [D. Lavenier]
- Briefing in Bioinformatics [D. Lavenier]
- Computers and Electronics in Agriculture [A. Mucherino]
- IEEE Transactions on Reconfigurable Technology and Systems [D. Lavenier]
- Journal of Biomedical and Health Informatics [D. Lavenier]
- Plos One [D. Lavenier]
- Journal of Computational and Applied Mathematics [A. Mucherino]
- International Transactions in Operational Research [A. Mucherino]
- Nature Scientific Reports [F. Legeai]
- Nucleic Acids Research [D. Lavenier]

10.1.4. Invited talks

- P. Peterlongo, *Reference-free NGS data analysis*, Litis, Rouen, March 2015.
- A. Mucherino, *Distance Geometry and Discretization Orders*, University of Aveiro, Portugal, May 2015.
- P. Peterlongo, *Mapping reads on de Bruijn graphs*, Bordeaux, Labri team, May 2015.
- A. Mucherino, *The several applications of the Distance Geometry*, University of Florianopolis, Brazil, June 2015.
- C. Lemaitre, *Reference-free detection of genomic variants: from SNPs to inversions*, "ABS4NGS", Institut Curie, Paris, France, June 2015.
- P. Peterlongo, *de Bruijn Graph usage and limitations*, Workshop on the future of algorithmic computational biology, Bertinoro, Italy, June 2015.
- D. Lavenier, *Genomic Data Processing*, "Journée Thématique GDR SoC-SiP", France, Nov. 2015.

- P. Peterlongo, *Prédiction de variants sans (ou avec) génome de référence*, MIA team, Toulouse, Nov. 2015.
- D. Lavenier, *Hybrid assembly based on long reads*, "Journée Scientifique Génomique et Bio-Informatique", France, on Dec. 2015.

10.1.5. Scientific expertise

- Expert for the MEI (International Expertise Mission), French Research Ministry [D. Lavenier]
- Member of the Scientific Council of BioGenOuest [D. Lavenier]
- Member of the Scientific Council of the Computational Biology Institute of Montpellier [D. Lavenier]

10.1.6. Research and teaching administration

- Member of « Conseil pédagogique de l'ISTIC » [R. Andonov]
- Responsible of the 3rd Year Computer Science BSc at ISTIC [R. Andonov]
- Member of the local Inria Rennes CDT (Technologic Transfer Commission) [D. Lavenier]
- Member of the steering committee of the INRA BIPAA Platform (BioInformatics Platform for Agro-ecosystems Arthropods) [D. Lavenier]
- Member of the steering committee of The GenOuest Platform (Bioinformatics Platform of BioGenOuest) [D. Lavenier]
- Representative of the environmental axis of UMR IRISA [C. Lemaitre]
- Inria center referee of Scientific mediation [P. Peterlongo]
- Member of the redaction committee Ouest Inria [P. Peterlongo]
- INRA Engineer recruitment committee [C. Lemaitre]
- Bordeaux University Engineer recruitment committee [C. Lemaitre]
- INRA Engineer recruitment committee, STLO, Rennes [F. Legeai]
- Assistant Professor recruitment committee, University of Brest [D. Lavenier]
- Professor recruitment committee, Polytech Montpellier [D. Lavenier]
- Scientific Responsible for International Relationships at ISTIC [A. Mucherino]
- Member of "Commission Affaires Internationales" at University of Rennes 1 [A. Mucherino]
- Member of the ISA Phd grant attribution jury [P. Peterlongo]
- AGOS first secretary [P. Peterlongo]

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

- Licence : A. Mucherino, Java basis, 80h, L1, Univ. Rennes 1, France.
- Licence : C. Lemaitre, Statistics for biology, 16h, L3, Univ. Rennes 1, France
- Master : A. Mucherino, Operational Research, 18h, M1, Univ. Rennes 1, France.
- Master : A. Mucherino, Introduction to Computational Systems and Networks, 42h, M1, Univ. Rennes 1, France. : A. Mucherino, Object Oriented Programming, 40h, M1, Univ. Rennes 1, France.
- Master : A. Mucherino, P. Peterlongo and R. Andonov, Algorithms on Sequences and Structures, 36h, M2, Univ. Rennes 1, France.
- Master : A. Mucherino, Parallel Computing (in English), 18h, M1, Univ. Rennes 1, France.
- Master : C. Lemaitre, P. Peterlongo, Text algorithmics for Bioinformatics, 43h, M1, Univ. Rennes 1, France.
- Master : C. Lemaitre, Dynamical systems for biological networks, 20h, M2, Univ. Rennes 1, France.

- Master : P. Peterlongo, Experimental Bioinformactics, 12h, M1, ENS Rennes, France.
- Master : D. Lavenier, Genomic Processing Data, 24h, M2, ESIR, Rennes, France.

10.2.2. Supervision

- PhD defense : Mathilde Le Boudic-Jamin, Similarités et divergences, globales et locales, entre structures proteiques, Univ. Rennes 1., 14/12/2015, supervised by R. Andonov.
- PhD defense : F. Moreews, Concevoir et échanger des workflows d'analyse de données. Application aux traitements intensifs en bioinformatique, Univ. Rennes 1, 11/12/2015, supervised by D. Lavenier and S. Lagarrigue.
- PhD in progress : G. Benoit, New algorithms for comparative metagenomics, start: 11/2014, D. Lavenier and C. Lemaitre
- PhD in progress : A. Limasset, Algorithm for Genomics, start: 09/2014, D. Lavenier and P. Peterlongo
- PhD in progress : C. Guyomar, Bioinformatic tools and applications for metagenomics of bacterial communities associated to insects, start: 10/2015, C. Lemaitre and F. Legeai
- PhD in progress : C. Marchet, Nouvelles méthodologies pour l'assemblage de données de séquençage polymorphes, start: 10/2015, P. Peterlongo
- PhD in progress : P. Hoan Son, Data mining and bioinformatics, start: 01/2015, D. Lavenier and A. Termier.

10.2.3. Juries

- *President of Ph-D thesis jury*. Julien Boutte, University of Rennes 1 [D. Lavenier]; Mouhamadou Ba, INSA Rennes [D. Lavenier], Clovis Galiez, University of Rennes 1 [R. Andonov]
- *Member of Ph-D thesis juries*. François Moreews, University of Rennes [D. Lavenier]
- *Referee of Ph-D thesis*. Nguyen, Thuy Diem, Nanyang Technical University, Singapor [D. Lavenier]; Andrea Radulescu, University of Nantes [D. Lavenier], Deepesh Agarwal, University of Nice [R. Andonov]
- *Member of Ph-D thesis comitees*. J. Laniau, University of Rennes [A. Mucherino]; A. Radulescu, university of Nantes [P. Peterlongo]; L. Siegwald, University of Lille [P. Peterlongo]; A. Mas, University of Rennes [P. Peterlongo]; H. Lopez, University of Lyon [C. Lemaitre]; T. Cumer, University of Grenoble [C. Lemaitre]; D. Eoche-Bosy, University of Rennes [F. Legeai]; H. Boulain, University of Rennes [F. Legeai]; Y., University of Rennes [F. Legeai]; M. Mulot, University of Colmar [F. Legeai].

10.3. Popularization

- Animation of Bioinformatics Atelier at Data Science Symposium (IRISA's 40th anniversary) "De la bioinformatique aux tiques". [P. Peterlongo]
- Open House day for IRISA's 40th anniversary: genomic puzzle [All members of GenScale]
- Participation to the event "A la decouverte de la recherche". [P. Peterlongo]
- Popularization paper in BioFutur [38]

11. Bibliography

Major publications by the team in recent years

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

[2] 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

[3] E. DREZEN, G. RIZK, R. CHIKHI, C. DELTEL, C. LEMAITRE, P. PETERLONGO, D. LAVENIER. *GATB: Genome Assembly & Analysis Tool Box*, in "Bioinformatics", 2014, vol. 30, pp. 2959 - 2961 [*DOI :* 10.1093/BIOINFORMATICS/BTU406], https://hal.archives-ouvertes.fr/hal-01088571

[4] N. MAILLET, C. LEMAITRE, R. CHIKHI, D. LAVENIER, P. PETERLONGO. *Compareads: comparing huge metagenomic experiments*, in "RECOMB Comparative Genomics 2012", Niterói, Brazil, October 2012, https://hal.inria.fr/hal-00720951

[5] N. MALOD-DOGNIN, R. ANDONOV, N. YANEV. *Maximum Cliques in Protein Structure Comparison*, in "SEA 2010 9th International Symposium on Experimental Algorithms", Naples, Italy, P. FESTA (editor), Springer, May 2010, vol. 6049, pp. 106-117 [*DOI :* 10.1007/978-3-642-13193-6_10], https://hal.inria.fr/inria-00536700

[6] 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

[7] V. H. NGUYEN, D. LAVENIER. *PLAST: parallel local alignment search tool for database comparison*, in "Bmc Bioinformatics", October 2009, vol. 10, 24 p. , http://hal.inria.fr/inria-00425301

[8] 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

[9] G. RIZK, A. GOUIN, R. CHIKHI, C. LEMAITRE. *MindTheGap: integrated detection and assembly of short and long insertions*, in "Bioinformatics", December 2014, vol. 30, n° 24, pp. 3451 - 3457 [*DOI :* 10.1093/BIOINFORMATICS/BTU545], https://hal.inria.fr/hal-01081089

[10] 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

[11] 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

[12] M. LE BOUDIC-JAMIN. *Global and Local Similarities and Divergences of Protein Structures* , Universite Rennes 1, December 2015, https://hal.inria.fr/tel-01250540

[13] F. MOREEWS. *Design and share data analysis workflows. Application to bioinformatics intensivetreatments*, université de rennes 1, December 2015, https://hal.inria.fr/tel-01233191

Articles in International Peer-Reviewed Journals

[14] N. ABBAS, S. DERRIEN, S. RAJOPADHYE, P. QUINTON, A. CORNU, D. LAVENIER. *Combining execution pipelines to improve parallel implementation of HMMER on FPGA*, in "Microprocessors and Microsystems", October 2015, vol. 39, pp. 457-470 [*DOI :* 10.1016/J.MICPRO.2015.06.006], https://hal.inria.fr/hal-01235328

[15] R. ANDONOV, H. DJIDJEV, G. KLAU, M. BOUDIC-JAMIN, I. WOHLERS. *Automatic Classification of Protein Structure Using the Maximum Contact Map Overlap Metric*, in "Algorithms and Combinatorics", December 2015, vol. 8, n^{o} 4 [*DOI :* 10.3390/A8040850], https://hal.inria.fr/hal-01248543

[16] G. BENOIT, C. LEMAITRE, D. LAVENIER, E. DREZEN, T. DAYRIS, R. URICARU, G. RIZK. *Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph*, in "BMC Bioinformatics", September 2015, vol. 16, n^{o} 1 [*DOI :* 10.1186/S12859-015-0709-7], https://hal.inria.fr/hal-01214682

[17] A. CASSIOLI, B. BARDIAUX, G. BOUVIER, A. MUCHERINO, R. ALVES, L. LIBERTI, M. NILGES, C. LAVOR, T. E. MALLIAVIN. *An algorithm to enumerate all possible protein conformations verifying a set of distance constraints*, in "BMC Bioinformatics", January 2015, vol. 16, n^{o} 1, 23 p. , https://hal.inria.fr/hal-01199667

[18] G. CHAPUIS, M. LE BOUDIC-JAMIN, R. ANDONOV, H. DJIDJEV, D. LAVENIER. *Parallel Seed-Based Approach to Multiple Protein Structure Similarities Detection*, in "Scientific Programming", April 2015, vol. 2015 [*DOI :* 10.1155/2015/279715], https://hal.inria.fr/hal-01235331

[19] P. DUMAS, F. LEGEAI, C. LEMAITRE, E. SCAON, M. ORSUCCI, K. LABADIE, S. GIMENEZ, A. L. CLAMENS, H. HENRI, F. VAVRE, J.-M. AURY, P. FOURNIER, G. KERGOAT, E. D'ALENÇON. *Spodoptera frugiperda (Lepidoptera: Noctuidae) host-plant variants: two host strains or two distinct species?*, in "Genetica", 2015, vol. 143, n^{o} 3, pp. 305-316 [*DOI :* 10.1007/S10709-015-9829-2], https://hal.archives-ouvertes.fr/hal-01208780

[20] H. DJIDJEV, G. CHAPUIS, R. ANDONOV, S. THULASIDASAN, D. LAVENIER. *All-Pairs Shortest Path Algorithms for Planar Graph for GPU-Accelerated Clusters*, in "Journal of Parallel and Distributed Computing", November 2015, vol. 85, pp. 91-103 [*DOI :* 10.1016/J.JPDC.2015.06.008], https://hal.inria.fr/hal-01235348

[21] N. GLASER, A. GALLOT, F. LEGEAI, M. HARRY, L. KAISER, B. LE RU, P.-A. CALATAYUD, E. JACQUIN-JOLY. *Differential expression of the chemosensory transcriptome in two populations of the stemborer Sesamia nonagrioides*, in "Insect Biochemistry and Molecular Biology", September 2015, vol. 65, pp. 28-34, https://hal.inria.fr/hal-01240447

[22] A. GOUIN, F. LEGEAI, P. NOUHAUD, A. WHIBLEY, J.-C. SIMON, C. LEMAITRE. *Whole-genome resequencing of non-model organisms: lessons from unmapped reads*, in "Heredity", May 2015, vol. 114, pp. 494-501 [*DOI :* 10.1038/HDY.2014.85], https://hal.inria.fr/hal-01081094

[23] F. LEGEAI, T. DERRIEN. *Identification of long non-coding RNAs in insects genomes*, in "Current Opinion in Insect Science", February 2015, vol. 7, pp. 37 - 44 [*DOI :* 10.1016/J.COIS.2015.01.003], https://hal.inria.fr/hal-01240461

[24] V. LOUX, M. MARIADASSOU, S. ALMEIDA, H. CHIAPELLO, A. HAMMANI, J. BURATTI, A. GENDRAULT, V. BARBE, J.-M. AURY, S.-M. DEUTSCH, S. PARAYRE, M.-N. MADEC, V. CHUAT, G. JAN, P. PETER-

LONGO, V. AZEVEDO, Y. LE LOIR, H. FALENTIN. *Mutations and genomic islands can explain the strain dependency of sugar utilization in 21 strains of Propionibacterium freudenreichii*, in "BMC Genomics", December 2015, vol. 16, n° 1, 35 p. [*DOI :* 10.1186/s12864-015-1467-7], https://hal.inria.fr/hal-01142363

[25] F. MOREEWS, O. SALLOU, H. MÉNAGER, Y. LE BRAS, C. MONJEAUD, C. BLANCHET, O. COLLIN. *BioShaDock: a community driven bioinformatics shared Docker-based tools registry*, in "F1000Research", December 2015 [*DOI :* 10.12688/F1000RESEARCH.7536.1], https://hal.inria.fr/hal-01243520

International Conferences with Proceedings

[26] E. JOLY, E. POIVET, A. GALLOT, F. LEGEAI, N. MONTAGNÉ. *The chemosensory transcriptome of the cotton leafworm Spodoptera littoralis*, in "24. Annual Meeting of the European-Chemoreception-Research-Organization (ECRO)", NA, France, Chemical Senses, Oxford University Press, 2015, vol. 40, n° 3, https://hal.archives-ouvertes.fr/hal-01208791

[27] A. MUCHERINO. *A Pseudo de Bruijn Graph Representation for Discretization Orders for Distance Geometry*, in "Proceedings of IWBBIO15", Granada, Spain, F. ORTUNO, I. ROJAS (editors), Lectures Notes in Computer Science, May 2015, vol. 9043, pp. 514–523, https://hal.inria.fr/hal-01196707

[28] N. PRAJAPATI, W. RANASINGHE, V. K. TANDRAPATI, R. ANDONOV, H. DJIDJEV, S. RAJOPADHYE. *Energy Modeling and Optimization for Tiled Nested-Loop Codes* , in "Parallel and Distributed Processing Symposium Workshop (IPDPSW)", Hyderabad , India, Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2015 IEEE International, May 2015, pp. 888 - 895 [*DOI :* 10.1109/IPDPSW.2015.94], https://hal.inria.fr/hal-01250602

National Conferences with Proceedings

[29] M. LE BOUDIC-JAMIN, R. ANDONOV. *De novo detection of structure repeats in Proteins*, in "Journées Ouvertes de Biologie, Informatique et Mathématiques JOBIM", Clermont-Ferrand, France, Société Française de BioInformatique (SFBI), July 2015, https://hal.inria.fr/hal-01250541

Conferences without Proceedings

[30] C. BETTEMBOURG, O. DAMERON, A. BRETAUDEAU, F. LEGEAI. *AskOmics : Intégration et interrogation de réseaux de régulation génomique et post-génomique*, in "IN OVIVE (INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement)", Rennes, France, June 2015, 7 p. , https://hal.inria.fr/hal-01184903

[31] F. MOREEWS, O. SALLOU, Y. LE BRAS, G. MARIE, C. MONJEAUD, T. A. DARDE, O. COLLIN, C. BLANCHET. *A curated Domain centric shared Docker registry linked to the Galaxy toolshed*, in "Galaxy Community Conference 2015", Norwich, United Kingdom, July 2015, https://hal.inria.fr/hal-01160514

[32] A. MUCHERINO. *Discretization Orders and Distance Geometry*, in "16ème conférence ROADEF :", Marseille, France, Proceedings of ROADEF2015, February 2015, https://hal.inria.fr/hal-01196717

[33] A. MUCHERINO. *Optimal Discretization Orders for Distance Geometry: a Theoretical Standpoint*, in "LSSC2015 - Proceedings of Large Scale Scientific Computations", Sozopol, Bulgaria, Lecture Notes in Computer Science, June 2015, https://hal.inria.fr/hal-01196701

[34] I. PETROV, S. BRILLET, E. DREZEN, S. QUINIOU, L. ANTIN, P. DURAND, D. LAVENIER. *KLAST: fast and sensitive software to compare large genomic databanks on cloud*, in "World Congress in Computer Science, Computer Engineering, and Applied Computing", Las Vegas, United States, July 2015, https://hal.inria.fr/hal-01235339

Scientific Books (or Scientific Book chapters)

[35] *Distance Geometry and Applications*, December 2015, vol. 197, https://hal.inria.fr/hal-01202376

[36] D. S. GONÇALVES, J. NICOLAS, A. MUCHERINO, C. LAVOR. *Finding Optimal Discretization Orders for Molecular Distance Geometry by Answer Set Programming*, in "Studies in Computational Intelligence", S. FIDANOVA (editor), Recent Advances in Computational Optimization, Springer, July 2015, vol. 610, pp. 1-15, https://hal.inria.fr/hal-01196714

[37] J. NICOLAS, P. PETERLONGO, S. TEMPEL. *Finding and Characterizing Repeats in Plant Genomes*, in "Plant Bioinformatics: Methods and Protocols", D. EDWARDS (editor), Methods in Molecular Biology, Humana Press - Springer Science+Business Media, November 2015, n^{o} 1374, 365 p. [*DOI :* 10.1007/978-1-4939-3167-5_17], https://hal.inria.fr/hal-01228488

Scientific Popularization

[38] D. LAVENIER. *Assembler un génome sur un pico-ordinateur*, in "Biofutur", July 2015, vol. 34, n^{o} 363, https://hal.inria.fr/hal-01235355

Other Publications

[39] G. BENOIT, P. PETERLONGO, D. LAVENIER, C. LEMAITRE. *Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets*, July 2015, JOBIM 2015, Poster [*DOI :* 10.1093/BIOINFORMA-CS/BTU406], https://hal.inria.fr/hal-01180603

[40] A. BODRUG. *Test and Benchmarking of A New Scaffolding Methodology*, University of Rennes 1, July 2015, https://hal.inria.fr/hal-01251303

[41] A. GOUIN, A. BRETAUDEAU, E. D'ALENÇON, C. LEMAITRE, F. LEGEAI. *Improvement of the assembly of heterozygous genomes of non-model organisms*, October 2015, Genome Informatics, Poster, https://hal.inria.fr/hal-01231793

[42] A. GOUIN, A. BRETAUDEAU, K. LABADIE, J.-M. AURY, E. D'ALENÇON, C. LEMAITRE, F. LEGEAI. *Improvement of the assembly of heterozygous genomes of non-model organisms, a case study of the genomes of two Spodoptera frugiperda host strains*, June 2015, Arthropod Genomics 2015, Poster, https://hal.inria.fr/hal-01240443

[43] A. LIMASSET, P. PETERLONGO. *BGREAT: A De Bruijn graph read mapping tool*, July 2015, JOBIM 2015, Poster, https://hal.inria.fr/hal-01192857

[44] C. RIOU, C. LEMAITRE, P. PETERLONGO. *VCF_creator: Mapping and VCF Creation features in DiscoSnp++*, July 2015, JOBIM 2015, Poster, https://hal.inria.fr/hal-01176492