IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Cachan

Université Rennes 1

Activity Report 2016

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:

3.1.2. - Data management, quering and storage
3.1.8. - Big data (production, storage, transfer)
3.3.2. - Data mining
3.3.3. - Big data analysis
7.1. - Parallel and distributed algorithms
7.3. - Optimization

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, HDR]
Guillaume Rizk [Associate Researcher, ANR HydroGen]

Faculty Members

Roumen Andonov [Univ. Rennes I, Professor, HDR]
Antonio Mucherino [Univ. Rennes I, Associate Professor, until Aug 2016]

Engineers

Fabrice Legeai [Inra]
Laurent Bouri [CNRS]
Jennifer Del Giudice [Inria, from Mar 2016]
Patrick Durand [Inria]
Sebastien Letort [CNRS]
Ivaylo Petrov [Inria, until Jan 2016]
Chloe Riou [Inria, until Apr 2016]
Charles Deltel [Inria, Research engineer, 50% time dedicated to the GenScale project]

PhD Students

Gaetan Benoit [Inria, granted by Hydrogen ANR project]
Sebastien Francois [Univ. Rennes I]
Cervin Guyomar [Univ. Rennes I]
Antoine Limasset [Univ. Rennes I]
Camille Marchet [Univ. Rennes I]
Hoang Son Pham [Vietnam gov.]

Post-Doctoral Fellow

Warley Gramacho Da Silva [Tocantins University, Brazil, until Aug 2016]

Visiting Scientists

Tristan Braquelaire [Univ. Bordeaux I, from Mar 2016 until Apr 2016]
Guillaume Chapuis [LANL, Los Alamos, NM, USA, June 2016]
Hristo Djidjev [LANL, Los Alamos, NM, USA, June 2016]

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 (Single Nucleotide Polymorphism (SNP), insertion, deletion, 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 analysis of large metagenomic projects does not 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 Nuclear Magnetic Resonance (NMR) 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 these 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 [3], [4].

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 [10], [8]. 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, especially 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], [9]. 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 [5], [2].

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 [6].

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 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. 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 on 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.

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. A few bioinformatics treatments, such as the 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 helps in prescribing the right drug. The bioinformatics analysis is based on the detection of SNPs (Single Nucleotide Polymorphism) from a set of target genes.

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

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 aim 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

- **Colib'read Workshop, Nov 7-8 th, Institut Curie, Paris.** GenScale organized a two-day workshop to present the main results of the Colib'read ANR (2013-2016, Coordinator P. Peterlongo) to the scientific community.
- **GATB Programming days.** In 2016, GenScale organized two Genome Analysis Toolbox (GATB) trainings days in Rennes (June 15 th) and Paris (Nov. 9 th). Each event gathered 15 persons who learned how to use the GATB library to design efficient NGS tools.

6. New Software and Platforms

6.1. AskOmics

KEYWORDS: RDF - SPARQL - Querying - Graph
FUNCTIONAL DESCRIPTION

AskOmics allows to load heterogeneous bioinformatics data (formatted as tabular files) into a Triple Store system using a user-friendly web interface. AskOmics also provides an intuitive graph-based user interface supporting the creation of complex queries that currently require hours of manual searches across tens of spreadsheet files. The elements of interest selected in the graph are then automatically converted into a SPARQL query that is executed on the users' data.

- Authors: Charles Bettembourg, Yvanne Chaussin, Anthony Bretaudeau, Olivier Filangi, Fabrice Legeai and Olivier Dameron
- Partners: CNRS - INRA - Université de Rennes 1
- Contact: Fabrice Legeai
- URL: https://github.com/askomics/askomics

6.2. BBhash

Basic binary representation of successive hash
KEYWORDS: C++ - Indexation - Data structures
FUNCTIONAL DESCRIPTION

BBHash is a simple library for building minimal perfect hash function. Given a set of N input keys, it will compute a bijective function that will associate to each key an integer between 1 and N. This then allows to create an indexed array that will hold some data for each key. It is designed to handle large scale datasets (hundred billion and more elements). The function itself is just a little bit larger than other state-of-the-art libraries, it takes approximately 3 bits / elements (compared to 2.62 bits/elem for the emphf lib), but construction is faster and does not require additional memory.

- Participants: Guillaume Rizk, Pierre Peterlongo, Rayan Chikhi and Antoine Limasset
- Contact: Guillaume Rizk
- URL: https://github.com/rizkg/BBHash

6.3. BCALM 2

KEYWORDS: Bioinformatics - NGS - Genomics - Metagenomics - De Bruijn graphs
SCIENTIFIC DESCRIPTION

BCALM 2 is a bioinformatics tool for constructing the compacted de Bruijn graph from sequencing data. It is a parallel algorithm that distributes the input based on a minimizer hashing technique, allowing for good balance of memory usage throughout its execution. It is able to compact very large datasets, such as spruce or pine genome raw reads in less than 2 days and 40 GB of memory on a single machine.
FUNCTIONAL DESCRIPTION

BCALM 2 is an open-source tool for dealing with DNA sequencing data. It constructs a compacted representation of the de Bruijn graph. Such a graph is useful for many types of analyses, i.e. de novo assembly, de novo variant detection, transcriptomics, etc. The software is written in C++ and makes extensive use of the GATB library.

- Participants: Rayan Chikhi, Antoine Limasset and Paul Medvedev
- Contact: Rayan Chikhi
- URL: https://github.com/GATB/bcalm

6.4. BGREAT

De bruijn graph read alignment tool
KEYWORDS: Short reads - Genome assembling
FUNCTIONAL DESCRIPTION

Mapping genomic extracts (reads) on genomic references is a central and necessary task in most genomic studies. But reference sequences are mainly extracted from assembly graphs through an inexact process that both creates chimeras and losses biological pieces of information. This motivates the need of mapping sequences on references represented by graphs. BGREAT is conceived to map reads on de Bruijn graph, a widely used graph in genome assembly.

- Participants: Pierre Peterlongo and Antoine Limasset
- Contact: Pierre Peterlongo
- URL: https://github.com/Malfoy/BGREAT

6.5. GATB-Core

Genome Assembly and Analysis Tool Box
KEYWORDS: Bioinformatics - NGS - Genomics - Genome assembling

FUNCTIONAL DESCRIPTION

The GATB-Core library 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 processor (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-Core library is written in C++.

- Participants: Dominique Lavenier, Guillaume Rizk, Pierre Peterlongo, Charles Deltel, Patrick Durand and Claire Lemaitre
- Contact: Dominique Lavenier
- URL: http://gatb.inria.fr/

6.6. GATB-Core Tutorial

Online GATB-Core tutorial
KEYWORD: Bioinformatics
FUNCTIONAL DESCRIPTION

"GATB-Core tutorial" is an interactive learning tool that aims at learning software development relying on the bioinformatics toolkit GATB-Core without the need of installing it, its dependencies and a C++ compiler. The tutorial relies on a client-server system. The client is simply a web browser running a full-featured C++ code editor. In turns, it is embedded in templates for the purpose of displaying various lessons. The server side is a Linux-based VM capable of compiling and running "online" any C++ code snippets using GATB-Core. That VM is deployed on Inria's AllGo SaaS platform.

- Participant: Patrick Durand
- Contact: Patrick Durand
- URL: http://gatb-core.gforge.inria.fr/training/

6.7. MindTheGap

KEYWORDS: Bioinformatics - NGS - Genome assembling
FUNCTIONAL DESCRIPTION

MindTheGap performs detection and assembly of DNA insertion variants in NGS read datasets with respect to a reference genome. It is designed to call insertions of any size, whether they are novel or duplicated, homozygous or heterozygous in the donor genome. The main algorithmic improvement of version 2.0.0 is to detect additionnal variants, such as SNPs and deletions. This feature improves the sensitivity of the insertion detection algorithm for insertions that are located near these other variants. Additionnally, 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.

- Participants: Claire Lemaitre and Guillaume Rizk
- Contact: Claire Lemaitre
- URL: https://gatb.inria.fr/software/mind-the-gap/

6.8. PLAST

Local alignment tool
KEYWORDS: Bioinformatics - Genomic sequence - Genomics
FUNCTIONAL DESCRIPTION

PLAST is a parallel alignment search tool for comparing large protein banks.

Sequence similarity searching is an important and challenging task in molecular biology and next-generation sequencing should further strengthen the need for faster algorithms to process such huge amount of data. At the same time, the internal architecture of current microprocessors is tending towards more parallelism, leading to the use of chips with two, four and more cores integrated on the same die. The main purpose of this work was to design an effective algorithm to fit with the parallel capabilities of modern microprocessors. A parallel algorithm for comparing large genomic banks and targeting middle-range computers has been developed and implemented in PLAST software. The algorithm exploits two key parallel features of existing and future microprocessors: the SIMD programming model (SSE instruction set) and the multithreading concept (multicore). Compared to multithreaded BLAST software, tests performed on an 8-processor server have shown speedup ranging from 3 to 6 with a similar level of accuracy.

- Participants: Dominique Lavenier, Erwan Drezen and Van Hoa Nguyen
- Contact: Dominique Lavenier
- URL: https://team.inria.fr/genscale/high-throughput-sequence-analysis/plast-intensive-sequence-comparison/

6.9. Simka

KEYWORDS: Comparative metagenomics - K-mer - Distance - Ecology
FUNCTIONAL DESCRIPTION

Simka is a comparative metagenomics method dedicated to NGS datasets. It computes a large collection of distances classically used in ecology to compare communities by approximating species counts by k-mer counts. The method scales to a large number of datasets thanks to an efficient and parallel kmer-counting strategy that processes all datasets simultaneoulsy.

- Participants: Gaetan Benoit, Claire Lemaitre, Pierre Peterlongo and Dominique Lavenier
- Contact: Gaetan Benoit
- URL: https://gatb.inria.fr/software/simka/

6.10. short read connector

KEYWORDS: Bioinformatics - Genomics - Metagenomics
SCIENTIFIC DESCRIPTION

Short read connector enables the comparisons of two read sets B and Q. For each read from Q it provides either:

The number of occurrences of each k-mers of the read in the set B (SRC_counter) or A list of reads from B that share enough k-mers with the tested read from B (SRC_linker)
FUNCTIONAL DESCRIPTION

This tool uses a data structure (BBHASH) adapted to the indexing of big data. Short Read Connector works on reads, which are sequencing data from high-throughput sequencers. Once the data is indexed, short read connector makes it possible either to find the similar reads in a dataset or to simply retrieve the approximate number of these similar reads.

- Participants: Pierre Peterlongo, Camille Marchet and Antoine Limasset
- Partner: UPMC
- Contact: Pierre Peterlongo
- URL: https://github.com/pierrepeterlongo/short_read_connector

7. New Results

7.1. HTS data processing

7.1.1. Providing end-user solutions, example from the Colib'read on galaxy project

Participants: Claire Lemaitre, Camille Marchet, Pierre Peterlongo.

Colib'read tools suite uses optimized reference-free algorithms for various analyses of NGS datasets, such as variant calling or read set comparisons. To facilitate data analysis and tools dissemination, we developed Galaxy tools and tool shed repositories. The galaxy package, facilitates the analyzis of raw NGS data for a broad range of life scientists [16].

7.1.2. Assembly of Streptococcus Bacteria

Participant: Dominique Lavenier.

With the microbiological and bacteriological group of the Rennes hospital, we design a new strategy to assemble the genomes of 40 *Streptococcus* bacteria. Each strain has been sequenced and independently assembled using different assembly tools. For a specific strain, a merge of the contigs is done using the MIX software. This step allows the number of contigs to be significantly reduced, resulting in a better final assembly compared to each individual assembly. The comparison with other known *Streptococcus* genomes indicates where phages are located in the genome [20].

7.1.3. Data-mining applied to GWAS

Participants: Pham Hoang Sun, Dominique Lavenier.

Identifying variant combination association with disease is a bioinformatics challenge. This problem can be solved by discriminative pattern mining that uses statistical functions to evaluate the significance of individual biological patterns. There is a wide range of such measures. However, selecting an appropriate measure as well as a suitable threshold in some specific practical situations is a difficult task. We propose to use the skypattern technique which allows combinations of measures to be used to evaluate the importance of variant combinations without having to select a given measure and a fixed threshold. Experiments on several real variant datasets demonstrate that the skypattern method effectively identifies the risk variant combinations related to diseases [28].

7.1.4. Variant detection in transcriptomic data

Participant: Camille Marchet.

We defined a method to identify, quantify and annotate SNPs (Single Nucleotide Polymorphisms) using RNA-seq reads only. Organisms with a poor quality or no reference genome can take benefit of this approach, as well as studies where not enough material is available for sequencing from one individual, where samples can be pooled. The method relies on motifs discovery and post-treatment in de Bruijn graphs built from the reads. It can be used for any species to annotate SNPs and predict their impact on proteins as well as test their association to a phenotype of interest. The approach has been validated using well known human RNA-seq data. Results have been compared with state of the art approaches for variant calling. We showed that the methods perform similarly in terms of precision and recall. Then we focused on the main target of the study, namely the non-model species. We finally validated experimentally the predictions of our method [18].

7.1.5. Faster de Bruijn graph compaction

Participant: Antoine Limasset.

We developed a new algorithm, called BCALM2, for the compaction of de Bruijn graphs. BCALM2 is a parallel algorithm based on minimizer repartition of sequences. This repartition allows the compaction of extremely large graphs with moderate memory usage and time. The compaction of a human sequencing graph can be done in 1 hour with only 3GB of memory and huge genomes, such as the pine and white spruce ones (more than 20Gbp each), can be handled using our approach on a regular server (2 days and 40GB of memory). Those results argue that BCALM2 is one order of magnitude more efficient than available approaches and can tackle the assembly bottleneck of constructing a compacted de Bruijn graph [14].

7.1.6. Scaffolding

Participants: Rumen Andonov, Sebastien François, Dominique Lavenier.

We developed a method for solving genome scaffolding as a problem of finding a long simple path in a graph defined by the contigs that satisfies additional constraints encoding the insert-size information. Then we solved the resulting mixed integer linear program to optimality using the Gurobi solver. We tested our algorithm on several chloroplast genomes and showed that it outperforms other widely-used assembly solvers by the accuracy of the results [25].

7.2. Sequence comparison

7.2.1. Metagenomics datasets comparison

Participants: Gaetan Benoit, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo.

We developed 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 2000 datasets) which aims at comparing worldwide sea water samples (ANR HydrGen project) [12].

7.2.2. Read similarity detection

Participants: Camille Marchet, Antoine Limasset, Pierre Peterlongo.

Retrieving similar reads inside or between read sets is a fundamental task either for algorithmic reasons or for analyses of biological data. This task is easy in small datasets, but becomes particularly hard when applied to millions or billions of reads. In [24] we used a straightforward indexing structure that scales to billions of elements. We proposed two direct applications in genomics and metagenomics. These applications consist in either approximating the number of similar reads between dataset(s) or to simply retrieve these similar reads. They can be applied on distinct read sets or on a read set against itself.

7.3. Parallelism

7.3.1. Processing-in-Memory

Participants: Charles Deltel, Dominique Lavenier.

The concept of PIM (Processor In Memory) aims to dispatch the computer power near the data. Together with the UPMEM company (http://www.upmem.com/), which is currently developing a DRAM memory enhanced with computing units, we investigate the parallelization of two bioinformatics algorithms for this new type of memory: sequence alignment and mapping [34] [33]. The first results show that blast-like algorithms or mapping algorithms can highly benefit from such memory and speed-up of more than 25 can be achieved [26].

7.3.2. GPU for graph algorithms

Participants: Rumen Andonov, Dominique Lavenier.

We describe three algorithms and their associated GPU implementations for two types of shortest path problems. These implementations target computations on graphs with up to millions of vertices and executions on GPU clusters. The first two algorithms solve the All-Pairs Shortest Path (APSP) problem. The first of these two algorithms allows computations on graphs with negative edges while the second trades this ability for better parallel scaling properties and improved memory access. The third algorithm solves the Single-Pair Shortest Path (SPSP) query problem. Our implementations efficiently exploit the computational power of 256 GPUs simultaneously. All shortest paths of a million vertex graph can be computed in 6 minutes and shortest path queries on the same graph are answered in a quarter of a millisecond. These implementations proved to be orders of magnitude faster than existing parallel approaches[30].

7.4. Data representation

7.4.1. Computational pan-genomics: status, promises and challenges

Participant: Pierre Peterlongo.

We took part to the Computational Pan-Genomics Consortium producing a "white paper" dedicated to computational pan-genomic. A pan-genome is a representation of the union of the genomes of closely related individuals (eg from a same species). Computational pan-genomics is a new sub-area of research in computational biology. In [19], we generalized existing definitions and we examined already available approaches to construct and use pan-genomes, discussed the potential benefits of future technologies and methodologies and reviewed open challenges from the vantage point of the above-mentioned biological disciplines.

7.4.2. Mapping reads on graphs

Participants: Pierre Peterlongo, Antoine Limasset.

Many published genome sequences remain in the state of a large set of contigs. Each contig describes the sequence found along some path of the assembly graph, however, the set of contigs does not record all the sequence information contained in that graph. Although many subsequent analyses can be performed with the set of contigs, one may ask whether mapping reads on the contigs is as informative as mapping them on the paths of the assembly graph.

In [17], we proposed a formal definition of mapping a sequence on a de Bruijn graph, we analysed the problem complexity, and we provided a practical solution. The proposed tool can map millions of reads per CPU hour on a de Bruijn graph built from a large set of human genomic reads. Results show that up to 22 % more reads can be mapped on the graph but not on the contig set.

7.5. Applications

7.5.1. Study of the rapeseed genome structure

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

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) being 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.

7.5.2. GATB Production Pipeline

Participants: Patrick Durand, Charles Deltel.

The entire set of libraries and tools related to the GATB Software have been introduced within a professional environment to support high-quality C++ developments. It relies on the use of technology platforms available at Inria: OpenStack and Jenkins. Considering the latter, we have setup more than 50 Jenkins tasks to automate the entire software development based on GATB: C++ code compiling and testing, documentation creation, packaging and preparation of official releases, mirroring on public Github repositories. Code compilation and tests are done on Linux and MacOSX VMs. https://ci.inria.fr/gatb-core/

7.5.3. Variant predictions in the pea genome

Participant: Pierre Peterlongo.

Progress in genetics and breeding in pea suffered from the limited availability of molecular resources. SNP markers that can be identified through affordable sequencing processes without the need for prior genome reduction or a reference genome allow the discovery of thousands of molecular markers.

We have been involved with IGEPP (Institut de Génétique, Environnement et Protection des Plantes, INRA) in the application of the discoSnp++ tool, discovering SNPs on HiSeq whole genome sequencing of four pea lines. Validation of a subset of predicted SNPs showed that almost all generated SNPs are highly designable and that most (95 %) deliver highly qualitative genotyping result [13].

7.5.4. Analysis of insect pest genomes

Participant: Fabrice Legeai.

Within a large international network of biologists, GenScale has contributed to various projects for identifying important components involved in the adaptation of major agricultural pests to their environment. We provided the assemblies, the annotations and the comparisons of various insects genomes [29]. Following specific agreement or policy, these results are available for browsing and consulting to a restricted consortium or a large community through the BioInformatics platform for Agro-ecosystems Arthropods (http://bipaa.genouest.org/is). In particular, this year our work helped to identify aphid genes involved in the adaptation to their favorite plant [15], or genes that are differentially expressed between leaf- and root-feeding phylloxera [21]. Furthermore, in order to help scientists to consult and cross genomics and postgenomics data, we are developping AskOmics, an integration and interrogation software for (linked) biological data, within a strong partnership, with Dyliss and GenOuest [36], [27].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Empowerd memory

Participants: Charles Deltel, Dominique Lavenier.

The UPMEM company is currently developing new memory devices with embedded computing power (http://www.upmem.com/). 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. EnginesOn start-up project

Participant: Jennifer Del Giudice.

EnginesOn is a start-up project based on life science digital data analysis (http://engineson.fr/). The origin of the project comes from a simple field observation: NGS technology is involved in numerous scientific studies. Deciphering the heterogeneous and voluminous data generated is a real challenge. People with the skills to analyze this type of data are scarce. EnginesOn focuses its first effort on health market with cancer diagnosis and personalized medecine. The start-up provides to physicians a virtual research laboratory with analysis workflows, compute infrastructure and data management that will lead to a simple, fast, reproducible diagnosis in a transparent fashion. EnginesOn also addresses the issue of big data management and storage. The project is entitled to the Fasttrack program since october 2016. Inria funds a 6-month technology transfer engineer in order to study the valorization and promote the GATB toolbox.

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 (http://www.rapsodyn.fr/). 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. 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. More precisely, we are in the process of setting up and evaluating a new method of predictions of small cancer-related mutations (such as SNPs and small insertions/deletions) from raw DNA sequencing data. The method relies on the use of k-mers and clustering of reads to call for mutations. Current prototype relies on Python programming language just for the purpose of evaluating the prediction quality of the software. However, final software is expected to use GATB library to highly increase the performance of the new tool.

9.1.2. Partnership with INRA in Rennes

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

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 of one INRA engineer (F. Legeai) and one PhD student (C. Guyomar).

9.2. National Initiatives

9.2.1. ANR

9.2.1.1. 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.

9.2.1.2. 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 do not require the availability of a reference genome. https://colibread.inria.fr/

9.2.1.3. Project HydroGen: Metagenomic applied to ocean life study

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

Coordinator: P. Peterlongo (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.

9.2.1.4. 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 butterfly 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, Sebastien Letort, Pierre Peterlongo.

Coordinator: N. Nesi (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.

9.2.3. Programs from research institutions

9.2.3.1. Inria ADT DiagCancer

Participants: Dominique Lavenier, Patrick Durand.

Since October 1st, 2016, Genscale has started a one-year Inria ADT called DiagCancer. It aims at: (1) including the DiscoSnp++ tool within the current data production pipeline at Pontchaillou Hospital (Rennes), (2) providing a new prediction tool applied to the calling of cancer related mutations from DNA sequencing data and (3) creating new analysis tools to facilitate the interpretation of results by end-users (biologists, doctors). The project is done in close collaboration with Haematology Service, CHU Pontchaillou, Rennes.

9.3. International Initiatives

9.3.1. Informal International Partners

- Free University of Brussels, Belgium: Genome assembly [P. Perterlongo, R. Andonov]
- IMECC, UNICAMP, Campinas, Brazil: Distance geometry problem [A. Mucherino]
- Los Alamos National Laboratory (LANL), Los Alamos: Graph structure, Parallelism, GPU [R. Andonov, D. Lavenier, G. Rizk]

9.4. International Research Visitors

9.4.1. Visits of International Scientists

- Visit of prof. Tomi Klein from Bar Ilan University (Israel). One week, december 2016. Collaboration for the application of approximate hash function to the TGS data analysis [P. Peterlongo]
- Visit of Hristo Djidjev from Los Alamos National Laboratory, June 2016. Graph algorithms for scaffolding problem, professeur invité, University of Rennnes 1, [R. Andonov]
- Visit of Guillaume Chapuis from Los Alamos National Laboratory, June 2016. Parallelism, GPU. [R. Andonov, D. Lavenier]

9.4.2. Internships

- Samyadeep Basu, BITS Pilani, India, May - July 2016. Development of a web server for assembling bacteria genomes [D. Lavenier, P. Durand, C. Deltel]

9.4.3. Visits to International Teams

9.4.3.1. Research Stays Abroad

- Visit of Guillaume Rizk to Los Alamos National Laboratory, USA, August - September 2016 (2 months). Efficient combinatorial optimization using quantum computing.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

- Workshop Colib'read. Scientific and practical organization [C. Lemaitre, C. Marchet, P. Peterlongo]

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

- WCO16, Gdansk, Poland (co-chair) [A. Mucherino]

10.1.2.2. Member of the Conference Program Committees

- BIBM 2016: IEEE International Conference on Bioinformatics and Biomedicine [D. Lavenier]
- BIOKDD 2016: 15th International Workshop on Data Mining in Bioinformatics [D. Lavenier]
- ACM PASC16 Conference : Platform for Advanced Scientific Computing [D. Lavenier]
- RECOMB 2016 : International Conference on Research in Computational Molecular Biology [D. Lavenier]
- ECCB 2016 : 15th European Conference on Computational Biology [P. Peterlongo]
- SeqBio 2016 : Bioinformatics multidisciplinary workshop [P. Peterlongo]
- WACEBI 2016 : Workshop on Accelerator-Enabled Algorithms and Applications in Bioinformatics [D. Lavenier]

10.1.2.3. Reviewer

- RECOMB 2017 [R. Andonov, C. Lemaitre]
- ECCB 2016 [P. Peterlongo]

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing Activities

- Algorithms for Molecular Biology [P. Peterlongo]
- Bioinformatics [D. Lavenier, P. Peterlongo]
- BMC Bioinformatics [P. Peterlongo]
- Drug Dicovery Today [D. Lavenier]
- Journal of Parallel and Distributed Computing [D. Lavenier]
- Plos One [D. Lavenier, R. Andonov]
- Journal of Biomedical and Health Informatics [D. Lavenier]
- Briefing in Bioinformatics [D. Lavenier]
- Theoretical Computer Science [P. Peterlongo]
- Fundamenta Informaticae, Integrated Computer-Aided Engineering (IOS Press) [A. Mucherino]
- COMPAG, Elsevier [A. Mucherino]

10.1.4. Invited Talks

- D. Lavenier, *Parallel Processing of Sequencing Data*, Conférence d'informatique en Parallélisme, Architecture et Système, Lorient, France, July 2016
- D. Lavenier *GATB: A Genomic Analysis Tool Box for designing parallel and low memory fingerprint bioinformatics software*, Pasteur Institut, Paris, France, Sep. 2016.
- D. Lavenier, *Low memory fingerprint data structure for genomics*, ENS Rennes, France, oct. 2016.
- D. Lavenier *GATB: A Genome Analysis Tool Box for designing parallel and low memory footprint bioinformatics software*, Workshop on Emerging Bioinformatics Applications for Microbial Ecogenomics, Brest, France, oct. 2016.
- C. Lemaitre, *Comparing numerous metagenomics datasets*, Laboratoire de Biométrie et Biologie Évolutive, Lyon, France, Nov. 2016.
- P. Peterlongo *De novo comparison of (large number of) metagenomic samples*, Metagenomics day, Billile, Lille, France June 2016.

- P. Peterlongo *Finding SNPs de novo from reads* , Bioadvection workshop, Napoli, Italy, June 2016.
- P. Peterlongo *De novo comparison of (large number of) metagenomic samples, What are the technical challenges? what can we expect from this?*, RCAM workshop - keynote speaker, The Hague, Netherlands, September 2016.
- P. Peterlongo *Multiple Comparative Metagenomics using Multiset k-mer Counting*, Pasteur Summer School 2016 In Metagenomics, September 2016.
- C. Lemaitre *Comparaison (massive) de (nombreux) metagénomes. Passons par les kmers pour passer à l'échelle*, Journée scientifique sur "le Microbiome" organisée par Biogenouest, Rennes, France, December 2016.
- F. Legeai *Les analyses bioinformatiques pour les données épigénomiques*, Atelier ChIP du réseau REacTION, Paris, France, December 2016.
- R. Andonov, *Global Optimization Methods for Genome Scaffolding and Completing Genome Assemblies*, Workshop on Graph Assembly Algorithms for omics data, Univ. Milano-Bicocca, Italy, November 18, 2016

10.1.5. Leadership within the Scientific Community

- P. Peterlongo. Animator of one of the scientific axes of the GDR BIM group of research.
- P. Peterlongo. Member of the SFBI board.

10.1.6. 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]

10.1.7. Research and Pedagogical Administration

- Member of the CoNRS, section 06, [D. Lavenier]
- 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 Agroecosystems Arthropods) [D. Lavenier]
- Member of the steering committee of The GenOuest Platform (Bioinformatics Platform of BioGenOuest) [D. Lavenier]
- Representative of the environnemental axis of UMR IRISA [C. Lemaitre]
- AGOS first secretary [P. Peterlongo]
- Organisation of the weekly seminar "Symbiose" [P. Peterlongo]
- Scientific Responsible for International Relationships at ISTIC [A. Mucherino]
- Member of "Commission Affaires Internationales" at University of Rennes 1 [A. Mucherino]
- In charge of the bachelor's degree in the computer science department of University of Rennes 1 (90 students) [R. Andonov]

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence : R. Andonov, Graph Algorithms, 60h, Univ. Rennes 1, France.

Licence : A. Mucherino, Java basis, 80h, L1, Univ. Rennes 1, France.

Master : A. Mucherino, R. Andonov, Operational Research, 78h, M1, Univ. Rennes 1, France.

Master : A. Mucherino, Introduction to Computational Systems and Networks, 42h, M1, Univ. Rennes 1, France.

Master : 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, Dynamical systems for biological networks, 20h, M2, Univ. Rennes 1, France.

Master : C. Lemaitre, P. Peterlongo, Algorithms on Sequences for Bioinformatics, 24h, M1, Univ. Rennes 1, France.

Master : P. Peterlongo, Experimental Bioinformactics, 12h, M1, ENS Rennes, France.

Master : D. Lavenier, Research training module, 24h, M1, Univ. Rennes 1, Rennes, France.

Master : R. Andonov, Advanced Algorithmics, 25h, Univ. Rennes 1, France.

Training : P. Durand, G. Rizk, GATB Programming Day, 8h (June 16th), Univ. Rennes 1, Rennes, France.

Training : P. Durand, G. Rizk, GATB Programming Day, 8h (May 9th), Institut Henri Poincaré, France.

E-learning

Online tutorial : GATB-Core Online Training tool. http://gatb-core.gforge.inria.fr/training/

discoSnp tutorial available from the discoSnp webpage: https://colibread.inria.fr/software/discosnp/

10.2.2. Supervision

HdR : Pierre Peterlongo, Lire les lectures : analyse de données de séquençage, Univ. Rennes 1, 25/01/2016, [11] https://hal.inria.fr/tel-01278275.

PhD in progress : G. Benoit, New algorithms for comparative metagenomics, 01/11/2014, D. Lavenier and C. Lemaitre.

PhD in progress : A. Limasset, Algorithm for Genomics, 09/2014, D. Lavenier and P. Peterlongo.

PhD in progress : C. Guyomar, Bioinformatic tools and applications for metagenomics of bacterial communities associated to insects, 01/10/2015, C. Lemaitre and F. Legeai.

PhD in progress : C. Marchet, Nouvelles méthodologies pour l'assemblage de données de séquençage polymorphes, 01/10/2015, P. Peterlongo.

PhD in progress : P. Hoan Son, Data mining and bioinformatics, 01/2015, D. Lavenier and A. Termier.

PhD in progress : S. François, Combinatorial Optimization Approaches for Bioinformatics, 01/10/2016, R. Andonov.

10.2.3. Juries

- *President of Ph-D thesis jury*. Phuong Do Viet, University of Montpellier [R. Andonov], Karel Brinda, University of Marne la Vallée [D. Lavenier]
- *Member of Ph-D thesis juries*. Joseph Lucas, University Pierre et Marie Curie [C. Lemaitre].
- *Referee of Ph-D thesis*. Cécile Monat, University of Montpellier [D. Lavenier], C. Vroland, University of Lille [P. Peterlongo].
- *Member of Ph-D thesis comitees*. L. Ishi Soares de Lima, University of Lyon [C. Lemaitre], Yoann Aigu, University of Rennes [F. Legeai], Hélène Boulain, University of Rennes [F. Legeai], Chunxiang Hao, University of Rennes [D. Lavenier], Alix Mas, University of Rennes [P. Peterlongo], Pierre Charrier, University of Nantes [P. Peterlongo], Cervin Guyomar, University of Rennes [P. Peterlongo], Lea Siegwald, University of Lille [P. Peterlongo].

10.3. Popularization

- Operation "A la découverte de la recherche" [P. Peterlongo]

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] G. BENOIT, P. PETERLONGO, M. MARIADASSOU, E. DREZEN, S. SCHBATH, D. LAVENIER, C. LEMAITRE. *Multiple comparative metagenomics using multiset k -mer counting*, in "PeerJ Computer Science", November 2016, vol. 2 [*DOI :* 10.7717/PEERJ-CS.94], https://hal.inria.fr/hal-01397150

[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^{o} 1, 22 p. [*DOI :* 10.1186/1748-7188-8-22], http://hal.inria.fr/hal-00868805

[4] 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

[5] 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

[6] 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

[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] 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^{o} 24, pp. 3451 - 3457 [*DOI :* 10.1093/BIOINFORMATICS/BTU545], https://hal.inria.fr/hal-01081089

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

[10] R. URICARU, G. RIZK, V. LACROIX, E. QUILLERY, O. PLANTARD, R. CHIKHI, C. LEMAITRE, P. PETERLONGO. *Reference-free detection of isolated SNPs*, in "Nucleic Acids Research", November 2014, pp. 1 - 12 [*DOI :* 10.1093/NAR/GKU1187], https://hal.inria.fr/hal-01083715

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] P. PETERLONGO. *Lire les lectures : analyse de données de séquençage*, Université rennes1, January 2016, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01278275

Articles in International Peer-Reviewed Journals

[12] G. BENOIT, P. PETERLONGO, M. MARIADASSOU, E. DREZEN, S. SCHBATH, D. LAVENIER, C. LEMAITRE. *Multiple comparative metagenomics using multiset k -mer counting*, in "PeerJ Computer Science", November 2016, vol. 2 [*DOI :* 10.7717/PEERJ-CS.94], https://hal.inria.fr/hal-01397150

[13] G. BOUTET, S. ALVES CARVALHO, M. FALQUE, P. PETERLONGO, E. LHUILLIER, O. BOUCHEZ, C. LAVAUD, M.-L. PILET-NAYEL, N. RIVIÈRE, A. BARANGER. *SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population*, in "BMC Genomics", December 2016, vol. 17, n^{o} 1, 121 p. [*DOI :* 10.1186/s12864-016-2447-2], https://hal.inria.fr/hal-01275696

[14] R. CHIKHI, A. LIMASSET, P. MEDVEDEV. *Compacting de Bruijn graphs from sequencing data quickly and in low memory*, in "Bioinformatics", November 2016, vol. 32, n^{o} 12, i201 - i208 [*DOI :* 10.1093/BIOINFORMATICS/BTW279], https://hal.archives-ouvertes.fr/hal-01395704

[15] I. EYRES, J. JAQUIÉRY, A. SUGIO, L. DUVAUX, K. GHARBI, J.-J. ZHOU, F. LEGEAI, M. NELSON, J.-C. SIMON, C. M. SMADJA, R. BUTLIN, J. FERRARI. *Differential gene expression according to race and host plant in the pea aphid*, in "Molecular Ecology", 2016, vol. 25, n^{o} 17, pp. 4197-4215 [*DOI :* 10.1111/MEC.13771], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01371828

[16] Y. LE BRAS, O. COLLIN, C. MONJEAUD, V. LACROIX, E. RIVALS, C. LEMAITRE, V. MIELE, G. SACOMOTO, C. MARCHET, B. CAZAUX, A. ZINE EL AABIDINE, L. SALMELA, S. ALVES-CARVALHO, A. ANDRIEUX, R. URICARU, P. PETERLONGO. *Colib'read on galaxy: a tools suite dedicated to biological information extraction from raw NGS reads*, in "GigaScience", February 2016, vol. 5, n^{o} 1 [*DOI :* 10.1186/s13742-015-0105-2], https://hal.inria.fr/hal-01280238

[17] A. LIMASSET, B. CAZAUX, E. RIVALS, P. PETERLONGO. *Read mapping on de Bruijn graphs*, in "BMC Bioinformatics", December 2016, vol. 17, n^{o} 1 [*DOI :* 10.1186/s12859-016-1103-9], https://hal.inria.fr/hal-01349636

[18] H. LOPEZ-MAESTRE, L. BRINZA, C. MARCHET, J. KIELBASSA, S. BASTIEN, M. BOUTIGNY, D. MONNIN, A. EL FILALI, C. M. CARARETO, C. VIEIRA, F. PICARD, N. KREMER, F. VAVRE, M.-F. SAGOT, V. LACROIX. *SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence*, in "Nucleic Acids Research", 2016 [*DOI :* 10.1093/NAR/GKW655], https://hal.inria.fr/hal-01352586

[19] T. MARSCHALL, M. MARZ, T. ABEEL, L. DIJKSTRA, B. E. DUTILH, A. GHAFFAARI, P. KERSEY, W. P. KLOOSTERMAN, V. MAKINEN, A. M. NOVAK, B. PATEN, D. PORUBSKY, E. RIVALS, C. ALKAN, J. A. BAAIJENS, P. I. W. D. BAKKER, V. BOEVA, R. J. P. BONNAL, F. CHIAROMONTE, R. CHIKHI, F. D. CICCARELLI, R. CIJVAT, E. DATEMA, C. M. V. DUIJN, E. E. EICHLER, C. ERNST, E. ESKIN, E. GARRISON, M. EL-KEBIR, G. W. KLAU, J. O. KORBEL, E.-W. LAMEIJER, B. LANGMEAD, M. MARTIN, P. MEDVEDEV, J. C. MU, P. NEERINCX, K. OUWENS, P. PETERLONGO, N. PISANTI, S. RAHMANN, B.

RAPHAEL, K. REINERT, D. D. RIDDER, J. D. RIDDER, M. SCHLESNER, O. SCHULZ-TRIEGLAFF, A. D. SANDERS, S. SHEIKHIZADEH, C. SHNEIDER, S. SMIT, D. VALENZUELA, J. WANG, L. WESSELS, Y. ZHANG, V. GURYEV, F. VANDIN, K. YE, A. SCHÖNHUTH. *Computational pan-genomics: status, promises and challenges*, in "Briefings in Bioinformatics", October 2016 [*DOI :* 10.1093/BIB/BBW089], https://hal.inria.fr/hal-01390478

[20] A. MEYGRET, V. PASCAL, S. MOULLEC, J. NACAZUME, Y. ADNANI, D. LAVENIER, S. KAYAL, A. FAILI. *Genome sequence of the uncommon Streptococcus pyogenes M/emm66 strain STAB13021, isolated from clonal clustered cases in French Brittany*, in "Genome Announcements", July 2016, vol. 4, n^{o} 4, 16 p. , e00689 [*DOI :* 10.1128/GENOMEA.00689-16], https://hal.inria.fr/hal-01385659

[21] C. RISPE, F. LEGEAI, D. PAPURA, A. BRETAUDEAU, S. HUDAVERDIAN, G. LE TRIONNAIRE, D. TAGU, J. JAQUIÉRY, F. DELMOTTE. *De novo transcriptome assembly of the grapevine phylloxera allows identification of genes differentially expressed between leaf- and root-feeding forms*, in "BMC Genomics", December 2016, vol. 17, n^{o} 1, 219 p. [*DOI :* 10.1186/s12864-016-2530-8], https://hal.inria.fr/hal-01286528

International Conferences with Proceedings

[22] S. FIDANOVA, O. ROEVA, A. MUCHERINO, K. KAPANOVA. *InterCriteria Analysis of Ant Algorithm with Environment Change for GPS Surveying Problem*, in "17th International Conference on Artificial Intelligence: Methodology , Systems, Applications (AIMSA16)", Varna, Bulgaria, C. DICHEV, G. AGRE (editors), Lecture Notes in Artificial Intelligence, Springer, September 2016, vol. 9883, pp. 271–278 [*DOI :* 10.1007/978-3-319-44748-3_26], https://hal.inria.fr/hal-01402412

[23] W. GRAMACHO, A. MUCHERINO, J.-H. LIN, C. LAVOR. *A New Approach to the Discretization of Multidimensional Scaling*, in "IEEE Conference Proceedings of FedCSIS16", Gandz, Poland, September 2016, https://hal.inria.fr/hal-01402390

[24] C. MARCHET, A. LIMASSET, L. BITTNER, P. PETERLONGO. *A resource-frugal probabilistic dictionary and applications in (meta)genomics*, in "Prageu Stringology Conference", Prague, Czech Republic, August 2016, https://hal.inria.fr/hal-01386744

Conferences without Proceedings

[25] S. FRANÇOIS, R. ANDONOV, H. DJIDJEV, D. LAVENIER. *Global Optimization Methods for Genome Scaffolding*, in "12th International Workshop on Constraint-Based Methods for Bioinformatics", Toulouse, France, September 2016, https://hal.inria.fr/hal-01385665

[26] D. LAVENIER, J.-F. ROY, D. FURODET. *DNA Mapping using Processor-in-Memory Architecture*, in "Workshop on Accelerator-Enabled Algorithms and Applications in Bioinformatics", Shenzhen, China, December 2016, https://hal.archives-ouvertes.fr/hal-01399997

[27] F. LEGEAI, C. BETTEMBOURG, A. BRETAUDEAU, Y. CHAUSSIN, O. DAMERON, D. TAGU. *BIPAA/Askomics, a new and easy approach for querying genomics and epigenomics elements in interaction*, in "XXVth International Congress of Entomology 2016", Orlando, Florida, United States, September 2016, https://hal.inria.fr/hal-01391080

[28] H.-S. PHAM, D. LAVENIER, A. TERMIER. *Identifying Genetic Variant Combinations using Skypatterns*, in "7th International Workshop on Biological Knowledge Discovery and Data Mining (Workshop BIOKDD

'16)", Porto, Portugal, DEXA, September 2016 [*DOI :* 10.1109/DEXA.2016.13], https://hal.inria.fr/hal-01385614

Scientific Books (or Scientific Book chapters)

[29] J. A. BRISSON, J. JAQUIÉRY, F. LEGEAI, G. L. TRIONNAIRE, D. TAGU. *Genomics of Phenotypic plastity in Aphids*, in "Management of Insect Pests to Agriculture", H. CZOSNEK, M. GHANIM (editors), Springer International Publishing, 2016, pp. 65-96 [*DOI :* 10.1007/978-3-319-24049-7_3], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01314988

[30] G. CHAPUIS, H. DJIDJEV, D. LAVENIER, R. ANDONOV. *GPU-accelerated shortest paths computations for planar graphs*, in "Advances in GPU Research and Practice", H. AZAD (editor), Elsevier, September 2016, 774 p. , https://hal.inria.fr/hal-01385634

[31] A. MUCHERINO, S. FIDANOVA, M. GANZHA. *Introducing the Environment in Ant Colony Optimization*, in "Studies in Computational Intelligence", Recent Advances in Computational Optimization, Springer, July 2016, vol. 655, pp. 147–158, https://hal.inria.fr/hal-01402423

Research Reports

[32] L. BOURI, D. LAVENIER. *Evaluation des logiciels d'assemblage utilisant des lectures longues*, Inria Rennes - Bretagne Atlantique, November 2016, n^{o} RT-0475, https://hal.inria.fr/hal-01282892

[33] D. LAVENIER, C. DELTEL, D. FURODET, J.-F. ROY. *BLAST on UPMEM*, Inria Rennes - Bretagne Atlantique, March 2016, n^{o} RR-8878, 20 p. , https://hal.archives-ouvertes.fr/hal-01294345

[34] D. LAVENIER, C. DELTEL, D. FURODET, J.-F. ROY. *MAPPING on UPMEM*, Inria, June 2016, n^{o} RR-8923, 17 p. , https://hal.archives-ouvertes.fr/hal-01327511

Other Publications

[35] G. BENOIT, P. PETERLONGO, M. MARIADASSOU, E. DREZEN, S. SCHBATH, D. LAVENIER, C. LEMAITRE. *Multiple Comparative Metagenomics using Multiset k-mer Counting*, April 2016, working paper or preprint, https://hal.inria.fr/hal-01300485

[36] A. EVRARD, C. BETTEMBOURG, M. M. JUBAULT, O. DAMERON, O. FILANGI, A. BRETAUDEAU, F. F. LEGEAI. *Integration and query of biological datasets with Semantic Web technologies: AskOmics*, June 2016, Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2016), Poster, https://hal.inria.fr/hal-01391087

[37] A. LIMASSET, C. MARCHET, P. PETERLONGO, L. BITTNER. *Minimal perfect hash functions in large scale bioinformatics Problem*, June 2016, JOBIM 2016, Poster, https://hal.archives-ouvertes.fr/hal-01341718

[38] C. MARCHET, A. LIMASSET, L. BITTNER, P. PETERLONGO. *A resource-frugal probabilistic dictionary and applications in (meta)genomics*, May 2016, working paper or preprint, https://hal.inria.fr/hal-01322440

[39] F. MOREEWS, D. LAVENIER. *Seamless Coarse Grained Parallelism Integration in Intensive Bioinformatics Workflows*, May 2016, working paper or preprint [*DOI :* 10.1145/2488551.2488588], https://hal.inria.fr/hal-00908842