IN PARTNERSHIP WITH:
CNRS

Université Rennes 1

Ecole normale supérieure de Cachan

Activity Report 2014

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

Keywords: Next Generation Sequencing, Genomics, Protein Structure, Parallelism, Graph

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

1. Members

Research Scientists

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

Faculty Members

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

Engineers

Fabrice Legeai [INRA]
François Moreews [INRA]
Susete Alves Carvalho [INRA]
Alexan Andrieux [Inria, until Sep 2014]
Laurent Bouri [CNRS, granted by France Genomique from Dec 2014]
Sébastien Brillet [Inria, from Mar 2014]
Erwan Drezen [Inria, granted by ANR GATB project]
Anaïs Gouin [Inria, granted by ANR ADA-SPODO project]
Ivaylo Petrov [Inria, granted by Brittany Region, from Oct 2014]
Chloé Riou [Inria, granted by ANR COLIB'READ project, from Oct 2014]
Charles Deltel [Inria, Research engineer, 50% time dedicated to the GenScale project]

PhD Students

Antoine Limasset [Univ. Rennes I, from Sep 2014]
Gaëtan Benoit [Inria, granted by ANR HYDROGEN project, from Nov 2014]
Mathilde Le Boudic-Jamin [Univ. Rennes I]

Post-Doctoral Fellows

Rodrigo Bentes Kato [granted by Brazilian Government, from Nov 2014]
Douglas Goncalves [CNRS, until Mar 2014]

Visiting Scientists

Ba Diep Nguyen [granted by CNRS Mastodons program, from Nov 2014]
Stephen Richards [INRA, from Jun 2014 until Jul 2014]

Administrative Assistants

Marie-Noëlle Georgeault [Inria, until Feb 2014]
Isobelle Kelly [Univ. Rennes I, from Feb until Oct 2014]
Marie Le Roïc [Univ. Rennes I, from Nov 2014]

Other

Guillaume Rizk [AlgoRizk, Research engineer, ANR GATB Project]

2. Overall Objectives

2.1. High throughput processing of genomic data

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

GenScale research activities cover the following domains:

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

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

3. Research Program

3.1. Introduction

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

- Data structure;
- Combinatorial optimization;
- Parallelism.

3.2. Data structure

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

3.3. Combinatorial optimization

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

3.4. Parallelism

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

4. Application Domains

4.1. Sequence comparison

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

4.2. Genome comparison

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

4.3. Protein comparison

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

5. New Software and Platforms

5.1. Next Generation Sequencing

Participants: Alexan Andrieux, Gaëtan Benoit, Charles Deltel, Erwan Drezen, Dominique Lavenier, Claire Lemaitre, Antoine Limasset, Pierre Peterlongo, Chloé Riou, Guillaume Rizk.

GATB: Genome 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 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 library is written in C++ and is available under the GNU Affero GPL License. [contact: D. Lavenier] https://gatb.inria.fr

Mapsembler: targetted assembly
The Mapsembler tool enables the micro assembly of one or several area(s) of interest. It takes as input one or more read set(s) and a one or more sequences fragments used as "starters" of each micro-assembly. This task provides a way to check the existence/absence of an area for which the user has an *a priori* interest. Moreover, for each extended "starter", the output is either a flat fasta sequence or a portion of the assembly graph. In this latter case, Mapsembler offers a visualization interface on which each graph (including the read coverage per read set) can be visualized, annotated, and manipulated. [contact: P. Peterlongo] http://colibread.inria.fr/mapsembler2/

Leon: NGS data compressor
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 encoding of a *C. elegans* reads set with 0.7 bits/base, outperforming state of the art reference-free methods. Leon is available under the GNU Affero GPL License. [contact: C. Lemaitre] https://gatb.inria.fr/software/leon/

Bloocoo: read corrector

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 contained in the GATB library, and inserts solid k-mers in a bloom-filter. The correction procedure is similar to state-of-the-art approaches. 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 [32]. [contact: C. Lemaitre] https://gatb.inria.fr/bloocoo-read-corrector/

MindTheGap: insertion variant detection
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: C. Lemaitre] http://mindthegap.genouest.org/

TakeABreak: de novo inversion variant discovery
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). TakeABreak is available under the GNU Affero GPL License. [contact: C. Lemaitre] http://colibread.inria.fr/software/takeabreak/

discoSnp: de novo SNP discovery
The discoSnp tool detects isolated SNPs given one, two or more raw read set(s) without using any reference genome. discoSnp ranks predictions and outputs quality and coverage per allele. Compared to finding isolated SNPs using a state-of-the-art assembly and mapping approach, discoSnp requires significantly less computational resources, shows similar precision and recall values, and highly ranked predictions are less likely to be false positives. [contact: P. Peterlongo] http://colibread.inria.fr/discosnp/

5.2. High throughput sequence comparisons

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

KLAST: bank-to-bank alignment search tool
KLAST 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. KLAST is a new optimized implementation of the PLAST algorithm to which several improvements have been made in 2014. KLAST is fully designed to compare query and subject comprised of large sets of DNA, RNA and protein sequences. It is significantly faster than original PLAST, while providing comparable sensitivity to BLAST and SSearch algorithms. KLAST contains a fully integrated data-filtering engine capable of selecting relevant hits with user-defined criteria (E-Value, identity, coverage, alignment length, etc.). Klast is developed with the Korilog Company and an academic version is now freely available for the scientific community [contact: D. Lavenier]. [34] https://koriscale.inria.fr/klast/

COMMET: de novo comparison of metagenomic datasets
Commet is an extension of the Comparead tool that proposes to compute similarity between set of raw non assembled (and usually non-assemblable with current state of the art assemblers) reads. Commet enables to factorize computations when n read sets have to be compared all together. Moreover, Commet proposes a new representation of sub-read sets that has the main advantages to save huge disk space and to enable efficient logical operations between sub-read sets. [contact: P. Peterlongo] https://colibread.inria.fr/software/commet/

5.3. 3D Protein structures

Participants: Douglas Goncalves, Antonio Mucherino.

MD-jeep version 0.2
MD-jeep is the result of a strong collaboration among Antonio Mucherino, Leo Liberti, Carlile Lavor and Nelson Maculan. Over the years, PhD and postdoc students under our supervision have also been contributing to this research topic. The new method for the computation of atomic coordinates in MD-jeep v.0.2 was developed in collaboration with Douglas Soares Gonçalves [13], who was a postdoc student in Rennes for one year [contact: A. Mucherino]. http://www.antoniomucherino.it/en/mdjeep.php

6. New Results

6.1. Highlights of the Year

discoSnp published in NAR. The publication presents a wide range of discoSnp applications that highlight the advantages and the drawbacks of predicting SNPs when no reference genomes are available. The publication witnesses the enthusiasm of users regarding both reference-free methods and the quality of the method. [20]

6.2. NGS methodology

Participants: Erwan Drezen, Anaïs Gouin, Dominique Lavenier, Claire Lemaitre, Antoine Limasset, Pierre Peterlongo, Guillaume Rizk.

Comparison of large sets of metagenomics data
We significantly extend the previous method (implemented in the Compaream tool) for computing similarity between sets of raw non assembled (and usually non-assemblable with current state of the art assemblers) reads. This enhancement of the method enables computations to be factorized when N read sets have to be compared all together. Moreover, the great advantage of this improvment is to save huge disk space and to enable efficient logical operations between metagenomic subset of reads. The Commet tool implements this optimized version.[25]

De novo SNP discovery
We developed a very efficient new way for detecting isolated SNPs given one, two or more raw read set(s) without using any reference genome. The implementation, called discoSnp, was applied to various datasets and applications. In particular, compared to finding isolated SNPs using a state-of-the-art assembly and mapping approach, our method requires significantly less computational resources, shows similar precision/recall values, and highly ranked predictions are less likely to be false positives. An experimental validation was conducted on an arthropod species (the tick *Ixodes ricinus*) on which de novo sequencing was performed. Among the predicted SNPs that were tested, 96% were successfully genotyped and truly exhibited polymorphism. [20]

De novo discovery of inversion breakpoints
A formal model has been proposed, together with an algorithm, for detecting inversion breakpoints without a reference genome, directly from raw NGS data. This model is characterized by a fixed size topological pattern in the de Bruijn Graph. We describe precisely the possible sources of false positives and false negatives and we additionally propose a sequence-based filter giving a good trade-off between precision and recall of the method. We implemented these ideas in a software called TakeABreak. Applied on simulated inversions in genomes of various complexity (from E. coli to a human chromosome dataset), the method provided promising results with a low memory footprint and a small computational time. [24]

Integrated detection and assembly of long insertion variants

We investigated a new method for the integrated detection and assembly of insertion variants from re-sequencing data. Contrary to other tools, it is designed to call insertions of any size, whether they are novel or duplicated, homozygous or heterozygous in the donor genome. We uses an efficient k-mer based method to detect insertion sites in a reference genome, and subsequently assemble them from the complete set of donor reads. The method is implemented in the tool MindTheGap and showed high recall and precision on simulated datasets of various genome complexities. When applied to real *C. elegans* and human datasets, MindTheGap detected and correctly assembled insertions longer than 1 kb, using at most 14 GB of memory. [19], [40]

Enhancement of de-Bruijn Graph data structure
The data structure holding the de-Bruijn Graph at the core of the GATB library has been improved through several new developments. First, its construction time has been greatly decreased thanks to the use of minimizers for kmer-counting, and efficient parallelization of various construction steps. Secondly, exploration of the graph has also been made faster through the possibility of parallel enumeration of nodes of interest, and through the use of a cache-coherent (blocked) bloom filter. Lastly, the structure itself has been extended to optionally allow for more information to be held, at a reasonable memory cost. A minimal perfect hash function allows to store additional data for each node, for example the coverage of each kmer. [11], [35], [36]

Chloroplast assembly
When sequencing plants, reads that correspond to the chloroplast genome are often over-represented. Filtering these reads based on k-mer counts allows specific assembly of the chloroplast to be directly performed. The small number of contigs can then be processed using advanced optimization tools to generate scaffolds. The approach has been partially tested on seqencing data from *Lactococcus lactis* to assemble plasmids of this bacteria. [12]

6.3. NGS applications

Participants: Susete Alves Carvalho, Rumen Andonov, Anaïs Gouin, Fabrice Legeai, Dominique Lavenier, Claire Lemaitre, Pierre Peterlongo, Ivaylo Petrov, Guillaume Rizk.

Identification of genomic regions of biological interest
The extraction and selection of 400 microsatellites among the large and fragmented *Acyrthosiphon pisum* genome led to the identification of a single 9cM region controlling the loss of sex in the pea aphid. The genotyping of these markers on geographically distant populations under divergent selection for reproductive strategies revealed a strong signature of selection in this genomic region, suggesting gene flow between populations with distinct reproductive modes. [15]

Transcriptome assembly
For this study, we incorporated various sources of RNA sequences from 454, Illumina and Sanger sequencing technologies obtained from more than 35 *S. frugiperda* developmental time-points and tissue samples and developed a custom pipeline to achieve their assembly. As a result, we provided a first valid transcriptome for *Spodoptera frugiperda*, a major agricultural pest. [16]

Catalogue of long non coding RNAs
We established a new bioinformatics pipeline for the detection of lncRNAs from RNA-Seq data, produced the first catalogue of aphids lncRNAs, and asserted for each lncRNA a classification of putative cis-interactions based on its genomic distance to neighboring mRNAs. These results allow the constitution of a broad gene regulation network of the aphid phenotypic plasticity at the embryo level. This workflow is available in Galaxy on the BioInformatics Platform for Arthopods of Agroecosystems (www.inra.fr/bipaa) and can be applied to any organism for which an annotated genome sequence and RNA-Seq data are provided.[23]

Identification and correction of genome mis-assemblies due to heterozygosity

Assembly tools are more and more efficient to reconstruct a genome from next-generation sequencing data but some problems remain. One of them corresponds to mis-assemblies due to heterozygosity (2 alleles instead of a consensus). Thus, we propose a strategy to detect and correct false duplications in assemblies based on several metrics: sequence similarity, matche length and average read coverage. Our method allows to decrease redundancy in the genome assembly, to improve the scaffolding and then to increase the N50 statistic by removal of one of the two alleles or joining of scaffolds by their extremities. This method was applied on the *Spodoptera frugiperda* genome.[39]

Questioning the classical re-sequencing analyses approach

Classical re-sequencing analyses are based on a first step of read mapping, then only mapped reads are taken into account in following analyses such as variant calling. We investigated the sources of unmapped reads in aphid re-sequencing data of 33 individuals, and we demonstrated that these reads contain valuable information that should not be discarded as usually done in such analyses. For instance, the analysis of the contigs obtained from assembling the unmapped reads led to recover some divergent genomic regions previously excluded from analysis and to discover putative novel sequences of *A. pisum* and its symbionts. We proposed strategies, based on assembly and re-mapping, to aid the capture and interpretation of this information.[14]

Application of discoSnp on pea data

The pea is a non-model organism with a large (4.5 GB) and complex genome which has not been sequenced yet. We compared, on the same set of low depth pea sequences, the SNPs generated by discoSnp with those obtained with a previous SNP discovery pipeline, and those generated using classical mapping approach combining Bowtie2 and GATK tools. [31]

6.4. HPC and parallelism

Participants: Rumen Andonov, Charles Deltel, Dominique Lavenier, François Moreews, Ivaylo Petrov.

Workflows

New tools are needed to enable the quick design and the intensive parallel execution of bioinformatics processes. Therefore, we propose a new Dataflow oriented workflow management system dedicated to intensive bioinformatics tasks. We worked on the interoperability of bioinformatics workflows using a model driven approach. Our results enable new import / export capabilities between multiple workflow management environments and incite to create a unique shared workflow model.[28]

Graph processing : the All-Pairs Shortest Paths problem

This research work anticipates the need of processing huge graphs that are results of intensive genomic sequence comparison (bank to bank processing). We proposed a new algorithm for solving the all-pairs shortest-path problem for planar graphs and graphs with small separators that exploits the massive on-chip parallelism available in today's Graphics Processing Units (GPUs). Our algorithm, based on the Floyd-War shall algorithm, has near optimal complexity in terms of the total number of operations, while its 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 the fastest known Dijkstra-based GPU implementation and a two-fold speedup over a parallel Dijkstra-based CPU implementation.[27]

Benchmark of Alignment Search Tools

Comparing sequences is a daily task in bioinformatics and many software try to fulfill this need by proposing fast execution times and accurate results. Introducing a new software in this field requires to compare it to recognized tools with the help of well defined metrics. A set of quality metrics is proposed that enables a systematic approach for comparing alignment tools. These metrics have been implemented in a dedicated software, allowing to produce textual and graphical benchmark artifacts. [21]

6.5. Protein Structure

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

The molecular distance geometry problem

The distance geometry is the problem of finding an embedding of a simple weighted undirected graph $G = (V, E, d)$ in a given dimension $K > 0$. Its most interesting application arises in biology, where the conformation of molecules such as proteins can be identified by embedding a graph (representing the molecular structure and some distance information) in dimension 3. Since some years, we are working on the discretization of the distance geometry. This year, the research developed in 4 main directions, that will be briefly detailed in the following paragraphs.

The majority of the work was performed on the so-called *discretization orders*, which are particular orders for the atoms of a molecule that allow for satisfying the discretization assumptions, i.e. they allow to discretize the search domain of the problem. Finding discretization orders is therefore an important pre-processing step for the solution of distance geometry problems. In fact, not only the identification of an atomic order allowing for the discretization is important, but also the identification of orders that are able to optimize some objectives that make the solution to the problem easier to perform. In this context, with both international and local partners, we worked on discretization orders that can be identified automatically in polynomial time [13], we worked on suitable orders for the protein side chains [10], and we studied some objectives to be optimized in discretization orders [38].

The algorithm that we mostly employ for the solution of distance geometry problems that can be discretized is the Branch & Prune (BP) algorithm. It recursively constructs the discretized search domain (a tree) and verifies the feasibility of the computed atomic positions. When all available distances are exact, all candidate positions for a given atom can be enumerated. This is however not possible in presence of interval distances, because a continuous subset of positions can actually be computed for the corresponding atoms. The focus of the work in [22] is on a new scheme for an adaptive generation of a discrete subset of candidate positions from this continuous subset. The generated candidate positions do not only satisfy the distances employed in the discretization process, but also additional distances that might be available (the so-called pruning distances).

Since the BP algorithm can loose in performance when dealing with large molecules containing several interval distances, we worked this year on a variation of the algorithm named BetaMDGP [29]. This is a work in collaboration with Korean researchers. The BetaMDGP algorithm is based on the concept of beta-complex, which is a geometric construct extracted from the quasi-triangulation derived from the Voronoi diagram of atoms.

From the theoretical side, we worked on two main directions. First, we proved that, in discretizable distance geometry problems where all available distances are exact, the total number of solutions is always a power of two. This is related to the fact that the discrete search space contains several symmetries [18]. Secondly, we tried to summarize in [37] the current issues for efficiently solving real-life instances of the distance geometry.

Finally, the work we performed during the last years, including another important results from other colleagues currently working on this topic, was summarized in an extensive survey on the discretization of the distance geometry [17].

Distance measure between Protein structure

We propose here a new distance measure for comparing two protein structures based on their contact map representations (CMO). This novel measure (max-CMO metric), satisfies all properties of a metric on the space of protein representations. Having a metric in that space allows to avoid pairwise comparisons on the entire database and thus to significantly accelerate exploring the protein space compared to non metric spaces. We show on a gold-standard classification benchmark sets that our exact k-nearest neighbor scheme classifies up to 95% and 99% of queries correctly. Our k-NN classification thus provides a promising approach for the automatic classification of protein structures based on contact map overlap. [26], [30]

Local similarity of protein structure

Finding similarities between protein structures is a main goal in molecular biology. Most of the existing tools preserve order and only find single alignments even when multiple similar regions exist. We propose a new seed-based approach that discovers multiple pairs of similar regions. Its computational complexity is polynomial and it comes with a quality guarantee that the returned alignments have both Root Mean Squared Deviations (coordinate-based as well as internal-distances based) lower than a given threshold, if such exists. We do not require the alignments to be order preserving, which makes our algorithm suitable for detecting similar domains when comparing multi-domain proteins. And because the search space for non-sequential alignments is much larger than for sequential ones, the computational burden is addressed by using both a coarse-grain level parallelism and a fine-grain level parallelism. [33]

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

7.1.1. Kalray Company : Parallelization of bioinformatics algorithm on the MPPA Platform

Participants: Charles Deltel, Dominique Lavenier.

The purpose was to investigate the performances of the Kalray MPPA architecture on scientific life science software. The collaboration started in 2013, and was aiming at implementing the PLAST software on the Kalray MPPA chip (256 cores). PLAST is a BLAST-like parallel implementation designed by GenScale. Experimentations have shown that for these kinds of applications that manage very huge volume of data, the MPPA chip memory capacity was a serious bottleneck.

7.1.2. Sofiproteol Company : Detection of SNPs in the Pea genome

Participants: Susete Alves Carvalho, Pierre Peterlongo.

The Peapol project is funded by Sofiproteol Company. Its mission is to develop the French vegetable oil and protein industry, open up new markets, and ensure an equal distribution of value among its members. The Peapol project counts two collaborators, Biogemma, and INRA, the latter working in collaboration with the GenScale team in charge of algorithmic research do detect SNPs in the pea genome.

7.2. Bilateral Grants with Industry

7.2.1. 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 research structure (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.

7.2.2. Rapsodyn project

Participants: Dominique Lavenier, Claire Lemaitre, 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.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. DGASP: Discrete Geometry Problem solve with ASP

Participants: Douglas Goncalves, Antonio Mucherino.

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

8.1.2. KoriKlast2: Intensive Sequence comparison

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

This is a collaborative project funded by Région Bretagne (18 months, from June 2014) with 3 partners: the Korilog Company, the bioinformatics computing center of Roscoff and the GenScale team. The purpose is (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.

8.1.3. Collaboration with IGDR (Insitute of Genetic and Development of Rennes)

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

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

8.1.4. Partnership with INRA

Participants: Susete Alves Carvalho, Anaïs Gouin, Dominique Lavenier, Fabrice Legeai, Claire Lemaitre, 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 three INRA engineers (F. Legeai, F. Moreews, S. Alves Carvalho). In particular, the collaboration with the IGEPP team includes several research projects in which Genscale is a formal partner: PEAPOL and SPECIAPHID projects.

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. 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.

8.2.1.2. Project SPECIAPHID: Speciation of pea aphids

Participants: Claire Lemaitre, Anaïs Gouin, Fabrice Legeai.

Coordinator: J-C. Simon (Inra)
Duration: 36 months (Jan. 2012 – Dec 2014)
Partners: IGEPP Inra Rennes, CBGP Inra Montpellier, BF2I Inra Lyon.

The SPECIAPHID project aims to understand the adaptation and speciation of pea aphids by re-sequencing and comparing the genomes of numerous aphid individuals. The role of Genscale is to apply and develop new methods to detect variation between re-sequenced genomes, and in particular complex variants such as structural ones.

8.2.1.3. 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: 39 months (Oct. 2012 – Dec 2015)
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.

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

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

Coordinator: P. Peterlongo (Inria, GenScale, Rennes)
Duration: 36 months (Mar. 2013 – Feb. 2016)
Partners: LIRMM Montpellier, Bamboo 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/

8.2.1.5. 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 an 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 used outside Genscale (LIRMM, Inria Bamboo team). http://gatb.inria.fr

8.2.1.6. 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.

8.2.2. PIA: Programme Investissement d'Avenir

8.2.2.1. RAPSODYN: Optimization of the rapeseed oil content and yield 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.

8.2.2.2. 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.

8.2.3. Programs from research institutions

8.2.3.1. Inria ADT Mapsembler

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

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

8.2.3.2. Mastodons CNRS Program SéPhHaDé: Computational Challenge of High Throughput Sequencing and Phenotyping in Life Science

Participants: Dominique Lavenier, Erwan Drezen, Ba Diep Nguyen.

Coordinator: E. Rivals (Lirmm, Montpellier)
Duration: 3 years (2012-2014)
Partners: Lirmm et Inria Montpellier, GenScale IRISA/Inria Rennes, Bamboo LIFL, Lille, INRA Montpellier , ISEM, IPMC Nice, CIRAD Montpellier, LSIS Aix Marseille, Tela Botanica Montpellier, UPMC Banyuls/Mer, CEA Evry, LITIS Rouen

This project deals with the management of huge volume of data generated (1) by the new sequencing technologies (2) by the collection of information for phenotyping living organisms. In 2014, GenScale has developed a methodology to compare metagenomic datasets to protein databanks.

8.3. International Initiatives

8.3.1. Inria International Partners

8.3.1.1. Informal International Partners

- Brazil
  - IMECC, UNICAMP, Campinas
  - COPPE, Federal University of Rio de Janeiro
  - University federal of Minas Gerais
- USA
  - Information Sciences Group (CCS-3), Los Alamos National Laboratory (LANL), Los Alamos.
  - Baylor College of Medicine, Houston
- China
  - StateKey Laboratory of Silkworm Genome Biology at the SouthWest University, Chongqing, China
- Vietnam
  - University of Cantho
- Europe
  - Bulgarian Academy of Science (BAS), Sofia, Bulgaria
  - The Genome Analysis Center, Norwich, UK
  - University of Sheffield, UK
  - University of York, UK

8.4. International Research Visitors

8.4.1. Visits of International Scientists

- Stephen Richards, Assistant Professor, Baylor College of Medicine, Houston, USA, June 2014. Stephen Richards is responsible for the sequencing and bioinformatics analysis of the genomes of arthropods. During his visit, he worked on the improvement of the pea aphid genome assembly.
- Ba Diep Nguyen, Assistant Professor, Cantho University, Vietnam Nov. 2014 to Jan. 2015 During his visit, Ba Diep Nguyen worked on the design of a new methodology for comparing metagenomic samples to protein databank.

8.4.2. Visits to International Teams

8.4.2.1. Research stays abroad

- Rumen Andonov, Professor, Information Sciences Group (CCS-3) from Los Alamos National Laboratory (LANL), Los Alamos, USA. Jan. 2014 to Aug. 2014. R. Andonov collaborates with LANL on various research projects related to solving hard combinatorial optimization problems on very large graphs and their applications in Bioinformatics. Two applications were on the focus of this cooperation during 2014: the scaffolding problem in NGS and structural classification of proteins.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. general chair, scientific chair

- Workshop on Computational Optimization (WCO14), Warsaw, Poland, Sep 7-11, 2014 [A. Mucherino, co-chair]

9.1.2. Scientific events selection

9.1.2.1. member of the conference program committee

- International Workshop on Algorithms for Computational Biology (ACB 2014) [D. Lavenier]
- IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2014) [D. Lavenier]
- European Conference on Computational Biology (ECCB'14) [D. Lavenier]
- International Conference on Field Programmable Logic (FPL 2014) [D. Lavenier]
- 4th Annual RECOMB Sattelite Workshop on Massively Parallel Sequencing (RECOMB-Seq 2014) [D. Lavenier, P. Peterlongo]
- International Conference on ReConFigurable Computing and FPGAs (ReConFig 2014) [D. Lavenier]
- IX Southern Programmable Logic Conference (SPL2014) [D. Lavenier]
- Many Faces of Distances (MFD14) [A. Mucherino]
- SeqBio 2014: Workshop on string algorithms [C. Lemaitre]

9.1.2.2. reviewer

- IWOCA 2014 [P. Peterlongo]

9.1.3. Journal

9.1.3.1. reviewer

- Advances in Bioinformatics [D. Lavenier]
- Algorithms for Molecular Biology [D. Lavenier]
- Bioinformatics [D. Lavenier]
- BMC Bioinformatics [D. Lavenier]
- BMC Genomics [D. Lavenier, F. Legeai]
- Briefing in Bioinformatics [D. Lavenier]
- Frontiers in Bioengineering and Biotechnology [P. Peterlongo]
- IEEE Transactions on Computational Biology and Bioinformatics [D. Lavenier]
- INS, CAMWA (Elsevier), ITOR (Wiley) , IJNS (World Scientific)[A. Mucherino]
- Integrative Biology [D. Lavenier]
- International Journal of Computer Science [P. Peterlongo]
- Journal of the Brazilian Computer Society [R. Andonov]
- Plos One [D. Lavenier]
- Nucleic Acids Research [D. Lavenier, P. Peterlongo]

9.1.4. Invited talks

- P. Peterlongo, Gen2Bio, Saint-Malo, France, Apr. 2014. Title: SNPs discovery.
- D. Lavenier, Los Alamos National Lab, New Mexico, USA, June 2014. Title: GATB: Genome Assembly and Analysis Tool Box.
- A. Mucherino, BAS, Sofia, Bulgaria, June 2014.
- D. Lavenier, Asprom seminar, Paris, France, Sept. 2014. Title: Parallel computation in genomics.
- P. Peterlongo, R-SYST, Inra, Rennes, France, Oct. 2014. Title: Comparing and combining multiple metagenomic datasets.

9.2. Administrative duties

- Member of the administrative council of ISTIC [R. Andonov]
- ANR Evaluation Committee Generic Call [R. Andonov]
- Member of the Scientific Council of Computational Biology Institute of Montpellier [D. Lavenier]
- Permanent expert for the MEI (International Expertise Mission), French Research Ministry [D. Lavenier]
- Member of the local Inria Rennes CDT (Technologic Transfer Commission) [D. Lavenier]
- Member of the 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]
- Member of the local Inria CORDIS committee for PhD grants [C. Lemaitre]
- 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]
- Recruitment committees: 1 professor [D. Lavenier], 2 assistant professors [C. Lemaitre, P. Peterlongo], 1 engineer (IE) [F. Legeai, chair]
- Scientific Responsible for International Relationships at ISTIC [A. Mucherino]
- Member of "Commission Affaires Internationales" at University of Rennes 1 [A. Mucherino]
- AGOS first secretary [P. Peterlongo]

9.3. Teaching - Supervision - Juries

9.3.1. Teaching

Licence : C. Lemaitre, Statistics for biology, 11h, L3, Univ. Rennes 1, France

Licence : D. Lavenier, Architecture and System, 30 h, L3, ENS Rennes, France.

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

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

Master : A. Mucherino and R. Andonov, Operational Research, 95h, M1, Univ. Rennes 1, France.

Master : R. Andonov, Advanced Algorithms, 15h, 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 : D. Lavenier, Intensive Computation of Genomic Data, 15 h, M1, ESEO Angers, France.

Master : D. Lavenier, Project, 12 h, M1, ENS Rennes, France.

Master : C. Lemaitre, P. Peterlongo, Text algorithmics for Bioinformatics, 43 h, M1, Univ. Rennes 1, France.

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

9.3.2. Supervision

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

PhD in progress : F. Moreews, Bioinformatics workflows, 01/2012, D. Lavenier and S. Lagarrigue

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

9.3.3. Juries

- *President of Ph-D thesis jury.* Anne Jeannin-Girardon, Université de Bretagne Occidentale, Brest [D. Lavenier]
- *Member of Ph-D thesis juries.* Germano Abud de Rezende, UNICAMP, Campinas, Sao Paulo, Brazil [A. Mucherino].
- *Referee of Ph-D thesis.* Guillaume Martin, Montpellier AgroSup, [D. Lavenier]
- *Member of Ph-D thesis committees.* J. Boutte, University of Rennes [D. Lavenier], P. Nouhaud, University of Rennes [C. Lemaitre and F. Legeai], C. Mercier, University of Grenoble [C. Lemaitre], S. Guizard, University of Tours [C. Lemaitre], A. Radulescu, university of Nantes [P. Peterlongo], H. Lopez, University of Lyon [C. Lemaitre], V Wucher, university of Rennes [F. Legeai], D. Eoche-Bosy, University of Rennes [F. Legeai], A. Marchant, University Paris XI [F. Legeai].

10. Bibliography

Major publications by the team in recent years

[1] R. Andonov, S. Balev, N. Yanev. *Protein Threading: From Mathematical Models to Parallel Implementations*, in "INFORMS Journal on Computing", 2004, vol. 16, n$^{\text{o}}$ 4, pp. 393-405 [*DOI :* 10.1287/ijoc.1040.0092], http://joc.journal.informs.org/content/16/4/393.abstract

[2] R. Andonov, N. Malod-Dognin, N. Yanev. *Maximum Contact Map Overlap Revisited*, in "Journal of Computational Biology", January 2011, vol. 18, n$^{\text{o}}$ 1, pp. 1-15 [*DOI :* 10.1089/cmb.2009.0196], http://hal.inria.fr/inria-00536624/en

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

[4] F. Legeai, G. Rizk, T. Walsh, O. Edwards, K. Gordon, D. Lavenier, N. Leterme, A. Mereau, J. Nicolas, D. Tagu, S. Jaubert-Possamai. *Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, Acyrthosiphon pisum*, in "BMC Genomics", 2010, vol. 11, n$^{\text{o}}$ 1, 281 p. [*DOI :* 10.1186/1471-2164-11-281], http://www.hal.inserm.fr/inserm-00482283

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

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

[7] P. Peterlongo, R. Chikhi. *Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer*, in "BMC Bioinformatics", March 2012, vol. 13, n$^{\text{o}}$ 48 [*DOI :* 10.1186/1471-2105-13-48], http://hal.inria.fr/hal-00675974

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

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

Publications of the year

Articles in International Peer-Reviewed Journals

[10] V. COSTA, A. MUCHERINO, C. LAVOR, A. CASSIOLI, L. M. CARVALHO, N. MACULAN. *Discretization Orders for Protein Side Chains*, in "Journal of Global Optimization", January 2014, vol. 60, n^{o} 2, pp. 333–349, https://hal.inria.fr/hal-01093052

[11] 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

[12] H. FALENTIN, D. NAQUIN, V. LOUX, F. BARLOY-HUBLER, P. LOUBIÈRE, S. NOUAILLE, D. LAVENIER, P. LE BOURGEOIS, P. FRANÇOIS, J. SCHRENZEL, D. HERNANDEZ, S. EVEN, Y. LE LOIR. *Genome Sequence of Lactococcus lactis subsp. lactis bv. diacetylactis LD61*, in "Genome Announcements", 2014, vol. 2, n^{o} 1, 1 p. [*DOI :* 10.1128/GENOMEA.01176-13], https://hal.archives-ouvertes.fr/hal-01019560

[13] D. GONÇALVES, A. MUCHERINO. *Discretization Orders and Efficient Computation of Cartesian Coordinates for Distance Geometry*, in "Optimization Letters", February 2014, vol. 8, n^{o} 7, pp. 2111–2125, https://hal.inria.fr/hal-01093049

[14] A. GOUIN, F. LEGEAI, P. NOUHAUD, A. WHIBLEY, J.-C. SIMON, C. LEMAITRE. *Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads*, in "Heredity", October 2014, 8 p. [*DOI :* 10.1038/HDY.2014.85], https://hal.inria.fr/hal-01081094

[15] J. JAQUIÉRY, S. STOECKEL, C. LAROSE, P. NOUHAUD, C. RISPE, L. MIEUZET, J. BONHOMME, F. MAHÉO, F. LEGEAI, J.-P. GAUTHIER, N. PRUNIER-LETERME, D. TAGU, J.-C. SIMON. *Genetic Control of Contagious Asexuality in the Pea Aphid*, in "PLoS Genetics", 2014, vol. 10, n^{o} 12, e1004838 [*DOI :* 10.1371/JOURNAL.PGEN.1004838], https://hal-univ-rennes1.archives-ouvertes.fr/hal-01093572

[16] F. LEGEAI, S. GIMENEZ, B. DUVIC, J.-M. ESCOUBAS, A.-S. GOSSELIN GRENET, F. BLANC, F. COUSSERANS, I. SÉNINET, A. BRETAUDEAU, D. MUTUEL, P.-A. GIRARD, C. MONSEMPES, G. MAGDELENAT, F. HILLIOU, R. FEYEREISEN, M. OGLIASTRO, A.-N. VOLKOFF, E. JACQUIN-JOLY, E. D'ALENÇON, N. NÈGRE, P. FOURNIER. *Establishment and analysis of a reference transcriptome for Spodoptera frugiperda*, in "BMC Genomics", 2014, vol. 15, n^{o} 1, 704 p. [*DOI :* 10.1186/1471-2164-15-704], https://hal.inria.fr/hal-01058982

[17] L. LIBERTI, C. LAVOR, N. MACULAN, A. MUCHERINO. *Euclidean Distance Geometry and Applications*, in "SIAM Review", February 2014, vol. 56, n^{o} 1, pp. 3–69, https://hal.inria.fr/hal-01093056

[18] L. LIBERTI, B. MASSON, J. LEE, C. LAVOR, A. MUCHERINO. *On the Number of Realizations of Certain Henneberg Graphs arising in Protein Conformation*, in "Discrete Applied Mathematics", March 2014, vol. 165, pp. 213–232, https://hal.inria.fr/hal-01093060

[19] 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

[20] 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

International Conferences with Proceedings

[21] E. DREZEN, D. LAVENIER. *Quality metrics for benchmarking sequences comparison tools*, in "BSB - 9th Brazilian Symposium on Bioinformatics", Belo Honrizonte, Brazil, S. CAMPOS (editor), Advances in Bioinformatics and Computational Biology Lecture Notes in Computer Science, Springer, October 2014, vol. 8826, pp. 144-153, https://hal.archives-ouvertes.fr/hal-01088595

[22] D. GONÇALVES, A. MUCHERINO, C. LAVOR. *An Adaptive Branching Scheme for the Branch & Prune Algorithm applied to Distance Geometry*, in "Federated Conference on Computer Science and Information Systems (FedCSIS14), Workshop on Computational Optimization (WCO14)", Warsaw, Poland, September 2014, pp. 463-469, https://hal.inria.fr/hal-01093063

[23] F. LEGEAI, T. DERRIEN, V. WUCHER, D. AUDREY, G. LE TRIONNAIRE, D. TAGU. *Long non--coding RNA in the pea aphid; identification and comparative expression in sexual and asexual embryos*, in "Arthropod Genomics Symposium", Urbana, United States, June 2014, https://hal.inria.fr/hal-01091304

[24] C. LEMAITRE, L. CIORTUZ, P. PETERLONGO. *Mapping-free and assembly-free discovery of inversion breakpoints from raw NGS reads*, in "Algorithms for Computational Biology", Tarragona, Spain, A.-H. DEDIU, C. MARTÍN-VIDE, B. TRUTHE (editors), July 2014, vol. 8542, pp. 119-130 [*DOI :* 10.1007/978-3-319-07953-0_10], https://hal.inria.fr/hal-01063157

[25] N. MAILLET, G. COLLET, T. VANNIER, D. LAVENIER, P. PETERLONGO. *COMMET: comparing and combining multiple metagenomic datasets*, in "IEEE BIBM 2014", Belfast, United Kingdom, November 2014, https://hal.inria.fr/hal-01080050

[26] I. WOHLERS, M. LE BOUDIC-JAMIN, H. DJIDJEV, G. W. KLAU, R. ANDONOV. *Exact Protein Structure Classification Using the Maximum Contact Map Overlap Metric*, in "AlCoB - 1st International Conference on Algorithms for Computational Biology", Tarragona, Spain, July 2014, pp. 262 - 273 [*DOI :* 10.1007/978-3-319-07953-0_21], https://hal.inria.fr/hal-01093803

Conferences without Proceedings

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

[28] F. MOREEWS, Y. LE BRAS, O. DAMERON, C. MONJEAUD, O. COLLIN. *Integrating GALAXY workflows in a metadata management environment*, in "Galaxy Community Conference", Baltimore, United States, July 2014, https://hal.inria.fr/hal-01093058

Scientific Books (or Scientific Book chapters)

[29] J. SEO, J.-K. KIM, J. RYU, C. LAVOR, A. MUCHERINO, D.-S. KIM. *BetaMDGP: Protein Structure Determination Algorithm Based on the Beta-complex*, in "Transactions on Computational Science XXII", M. GAVRILOVA, C. TAN (editors), Springer, February 2014, vol. Lecture Notes in Computer Science, n^{o} 8360, pp. 130–155, https://hal.inria.fr/hal-01093066

Research Reports

[30] I. WOHLERS, M. LE BOUDIC-JAMIN, H. DJIDJEV, G. W. KLAU, R. ANDONOV. *Exact Protein Structure Classification Using the Maximum Contact Map Overlap Metric*, Inria Rennes - Bretagne Atlantique and University of Rennes 1, France ; Genome Informatics, University of Duisburg-Essen, Germany ; Life Sciences, CWI, Science Park 123, 1098 XG Amsterdam, The Netherlands ; Los Alamos National Laboratory, Los Alamos, NM, USA, 2014, pp. 262 - 273 [*DOI :* 10.1007/978-3-319-07953-0_21], https://hal.inria.fr/hal-01093776

Other Publications

[31] S. ALVES CARVALHO, R. URICARU, J. DUARTE, C. LEMAITRE, N. RIVIÈRE, G. BOUTET, A. BARANGER, P. PETERLONGO. *Reference-free high-throughput SNP detection in pea: an example of discoSnp usage for a non-model complex genome*, September 2014, ECCB 2014, https://hal.inria.fr/hal-01091184

[32] G. BENOIT, D. LAVENIER, C. LEMAITRE, G. RIZK. *Bloocoo, a memory efficient read corrector*, September 2014, European Conference on Computational Biology (ECCB), https://hal.inria.fr/hal-01092960

[33] G. CHAPUIS, M. LE BOUDIC-JAMIN, R. ANDONOV, H. DJIDJEV, D. LAVENIER. *Parallel seed-based approach to multiple protein structure similarities detection* *, December 2014, https://hal.inria.fr/hal-01093809

[34] E. DREZEN, P. DURAND, D. LAVENIER. *KLAST: a new high-performance sequence similarity search tool*, April 2014, Bio-IT World Conference, https://hal.archives-ouvertes.fr/hal-01088629

[35] E. DREZEN, G. RIZK, R. CHIKHI, C. DELTEL, C. LEMAITRE, P. PETERLONGO, D. LAVENIER. *GATB: a software toolbox for genome assembly and analysis*, April 2014, Bio-IT World Conference, https://hal.archives-ouvertes.fr/hal-01088641

[36] E. DREZEN, G. RIZK, R. CHIKHI, C. DELTEL, C. LEMAITRE, P. PETERLONGO, D. LAVENIER. *GATB: Toolbox for developing efficient NGS software*, October 2014, BSB - 9th Brazilian Symposium on Bioinformatics, https://hal.archives-ouvertes.fr/hal-01088828

[37] D. GONÇALVES, A. MUCHERINO. *Challenges for Extending Discretizable Molecular Distance Geometry to Interval Data*, October 2014, Proceedings of Many Faces of Distances (MFD14), https://hal.inria.fr/hal-01093071

[38] D. GONÇALVES, J. NICOLAS, A. MUCHERINO. *Searching for Optimal Orders for Discretized Distance Geometry*, October 2014, Proceedings of Many Faces of Distances (MDF14), https://hal.inria.fr/hal-01093072

[39] A. GOUIN, A. BRETAUDEAU, C. LEMAITRE, F. LEGEAI. *Identification and correction of genome mis-assemblies due to heterozygosity*, September 2014, European Conference on Computational Biology (ECCB), https://hal.inria.fr/hal-01092959

[40] G. RIZK, A. GOUIN, R. CHIKHI, C. LEMAITRE. *MindTheGap: integrated detection and assembly of short and long insertions*, September 2014, European Conference on Computational Biology (ECCB), https://hal.inria.fr/hal-01087832