INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team sequoia

Algorithms for large-scale sequence analysis for molecular biology

Lille - Nord Europe

Theme : Computational Biology and Bioinformatics

Table of contents

SEQUOIA *is a joint project-team with LIFL (CNRS UMR 8022, Université Lille 1)*

1. Team

Research Scientist

Hélène Touzet [DR CNRS, Team leader, HdR]
Mathieu Giraud [CR CNRS]
Gregory Kucherov [DR CNRS, on leave for Poncelet Institute, Moscow, HdR]

Faculty Member

Jean-Stéphane Varré [MC, Université Lille 1, on leave at INRIA until September 2009, HdR]
Laurent Noé [MC, Université Lille 1]
Maude Pupin [MC, Université Lille 1, on leave at INRIA from September 2009]

Technical Staff

Antoine de Monte [Ingénieur, Université Lille 1, from July 2007]
Benjamin Grenier-Boley [Ingénieur Associé, INRIA, from September 2007 until August 2009]
Laurie Tonon [IJD, INRIA, from October 2009]

PhD Student

Ségolène Caboche [INRIA/Region fellowship, PhD in September 2009, now postdoctoral fellow]
Aude Darracq [MESR fellowship, from October 2007]
Arnaud Fontaine [PhD in March 2009, now postdoctoral fellow]
Marta Girdea [INRIA CORDI fellowship, from October 2007]
Azadeh Saffarian [MESR fellowship, from November 2007]
Tuan Tu Tran [INRIA CORDI fellowship, from September 2009]

Visiting Scientist

Peter Steffen [University of Bielefeld, March 2009, one month]

Administrative Assistant

Sandrine Catillon [INRIA]

2. Overall Objectives

2.1. Highlights of the year

- The NORINE database for nonribosomal peptides has been recognized by wwPDB as the international reference resource for peptides that are not encoded in genomes. The mission of the worldwide Protein Data Bank (wwPDB) is to maintain a single databank of macromolecular structural data, that is freely and publicly available to the global community.
- The RNAspace platform for annotation of noncoding RNAs has been awarded by the IBISA label. IBISA is a French consortium for evaluating and funding national technological platforms in life sciences.

2.2. Objectives

The main goal of SEQUOIA project-team is to define appropriate combinatorial models and efficient algorithms for large-scale sequence analysis in molecular biology. An emphasis is made on the annotation of non-coding regions in genomes – RNA genes and regulatory sequences – via comparative genomics methods. This task involves several complementary issues such as sequence comparison, prediction, analysis and manipulation of RNA secondary structures, identification and processing of regulatory sequences. Efficient algorithms and parallelism on high-performance computing architectures allow large-scale instances of such issues. Our aim is to tackle all those issues in an integrated fashion and to put together the developed software tools into a

common platform for annotation of non-coding regions. We also explore complementary problems of protein sequence analysis. Those include new approaches to protein sequence comparison on the one hand, and a system for storing and manipulating nonribosomal peptides on the other hand. A special attention is given to the development of robust software, its validation on biological data and to its availability from the software platform of the team and by other means. Most of research projects are carried out in collaboration with biologists.

3. Scientific Foundations

3.1. Introduction

From a historical perspective, research in bioinformatics started with string algorithms designed for the comparison of sequences. Bioinformatics became then more diversified, accompanying the emergence of new high-throughput technologies: DNA chips, mass spectrometry, and others. By analogy to the living cell itself, bioinformatics is now composed of a variety of dynamically interacting components forming a large network of knowledge: systems biology, proteomics, text mining, phylogeny, structural genomics,... Sequence analysis remains a central node in this interconnected network, and it is the heart of the SEQUOIA project. It is a common knowledge nowadays that the amount of sequence data available in public databanks (such as GenBank and others) grows at an exponential pace. The recent advent of new sequencing technologies, also called Next Generation Sequencing and deep sequencing, amplified this phenomenon. Sequencing a bacterial genome is now done routinely, at a very moderate cost. Even if the first draft human genome sequence was obtained only eight years ago, obtaining a genome sequence of an eukaryotic organism is currently becoming a routine and low-cost operation too. Next Generation Sequencing promises to revolutionize genomic and transcriptomic. It allows for a fast and low-cost massive acquisition of short genomic fragments and thus represent a remarkable tool for genome studies. It gives rise to new problems and gives new insight on old problems by revisiting them: accurate and efficient remapping/pre-assembling, fast and accurate search of non exact (but quality labelled) reads, and/or non species specific reads. To illustrate this, SOLiD technology enables error detection and correction, providing SNP detection accuracy at genome-wide scale even at sparse read coverage. Illumina announced in June 2009 a *personal genome sequencing service* offering a sequencing of an individual genome for only $48,000. This also opens the way to a variety of applications: contamination and vector detection, fast and accurate species detection in metagenomics, themselves having great potential in animal/human epidemic detection,...As a result, sequence analysis and sequence processing receive now a renewed attention [57].

The second incentive for sequence analysis is the progress in the elucidation of mechanisms of genome functioning. Molecular biology is a rapidly evolving science. Originally, sequence analysis was mostly driven by the scheme of the central dogma in its simplest formulation: information is contained in DNA, then it is transcribed into messenger RNA and finally translated into proteins. New pieces of information that shed a new light on the central dogma are now available. First, it is now widely recognized that the role of *noncoding RNA genes* has been largely underestimated until the late 90's. Following miRNAs and snoRNAs, many new families of those genes have been discovered recently: piRNAs, tasiRNAs, ... RNA genes are now known to play an important role in many cellular processes – protein synthesis, regulation. Furthermore, recent observations derived from tiling arrays or deep-sequencing technologies show that a large part of the transcriptional output of eukaryotic genomes does not appear to encode proteins.

Another biological phenomenon supplementing the central dogma occurs at the protein level. Translation of RNA is not the only way the proteins are synthesized in the cell: some peptides (typically in bacteria and fungi) result from a *nonribosomal synthesis* performed by a separate cell machinery. As the name suggests, it is an alternative pathway that allows for the production of polypeptides that are not encoded in the genome, and that are produced without ribosome but with other enzymatic complexes called nonribosomal synthetases (NRPSs). This biosynthesis has been described for the first time in the 70's [47]. For the last decade, the interest in nonribosomal peptides and their synthetases has considerably increased, as witnessed by the growing number

of publications in this field. These peptides are or can be used in many existing or potential biotechnological and pharmaceutical applications (e.g. anti-tumors, antibiotics, immuno-modulators).

Lastly, computer hardware is also evolving with the advent of *massively multicore processors*. For a few years, issues with heat dissipation prevent the processors from having higher frequencies. The thermal density of some processors approaches the one of the surface of the sun [44]. One of the answers to maintain the Moore's Law is the usage of parallel processing. Grid environments provide tools for effective implementation of coarse grain parallelization. Recently, another kind of hardware has attracted interest: multicore processors. Graphic processing units (GPUs) are a first step towards massively multicore processors. They allow everyone to have some teraflops of cheap computing power in its personal computer. High-end GPUs, for less than $500, embed far more arithmetic units than a CPU of the same price. Recent trends blur the line between such GPUs and CPUs. Moreover, libraries like CUDA (released in 2007) and OpenCL (specified in December 2008) facilitate the use of those units for general purpose computation. We believe that this new era in hardware architecture will bring new opportunities in large scale sequence analysis. For example, recent parallelizations on GPUs for sequence analysis problems achieve speedups between $10\times$ and $100\times$ compared to a serialized one-core version.

All above-mentioned biological phenomena together with big volumes of new sequence data and new hardware provide a number of new challenges to bioinformatics, both on modeling the underlying biological mechanisms and on efficiently treating the data.

3.2. Sequence processing and Next Generation Sequencing

Sequence algorithms is an established research subject of the team. We have been working on spaced seed techniques for several years and made several contributions, of which one of the most important is the concept of *subset seeds* [7] [45], [46]. The whole technique is implemented and made available in the YASS software for DNA sequence alignment together with the tools implemented to design seeds [10] (see Section 4.1). YASS has been used by many researchers and cited in several papers, some of them are mentioned on the YASS website (http://bioinfo.lifl.fr/yass/). We have also shown that the proposed seed model was perfectly adapted for protein search [59][20], and we developed such approach in massivelly parallel processing units [55]. We consider that we gained a good expertise in this area, supported by our theoretical work on algorithmic techniques and data structures.

Members of the team were among the first to work on advanced seeding methods for DNA sequence search [53], and joined, at early stages, the study of *spaced seed* design, started in [38], [49]. The techniques and tools we proposed have been used by the community (see Section 4.1) and cited in a number of papers and surveys (e.g. [37]). Main research groups in the world working on these topics are: Bioinformatics group at Waterloo University (B. Ma, H. Yao, D. Brown, M. Li, B. Brejova, T. Vinar), Departments of Computer Science and Engineering and Genetics at Washington University in St. Louis (J. Buhler, Y. Sun), Department of Computer Science at University of Western Ontario (K. Zhang and L. Ilie), Department of Mathematics at Singapore University (L. Zhang, K. P. Choi, F. Zeng and Y. Kong), Department of Computer Science and Information Engineering at Taiwan University (K.-M. Chao), Department of Computer Science at Ben-Gurion University of the Negev (M. Farach-Colton, G. Landau, S.C. Sahinalp, and D. Tsur), Departments of Biology and Computer Science at Boston University (G. Benson and D.Y. Mak), Center for Bioinformatics and Computational Biology at University of Maryland (L. Zhou and L. Florea).

3.3. Comparative analysis and Noncoding RNAs

Noncoding RNA analysis is another fundamental topic for SEQUOIA. Our first publications on this subject date back to 2003. First, we proposed a new method for RNA structure inference and implemented it in a program called CARNAC. CARNAC has been evaluated as one of the best available software tools for comparative RNA structure prediction in an independent comparative survey [43]. It also gave rise to a thorough presentation in a recent survey paper [50], and has been the subject of an invited book chapter in a general-purpose bioinformatics manual [62]. Since its first release, the CARNAC software has been regularly updated.

Second, we worked on theoretical models for RNA comparison. Comparison of RNA structures should take into account several levels of information corresponding to hierarchical RNA folding: sequence, secondary structure, tertiary interactions, involving complex combinatorial objects. RNA structures are usually modelled by ordered labeled trees or graphs, such as arc-annotated sequences. Our work led to substantial advances on tree edit distance algorithms [4] [63], [54], tree models [19], [18] and comparison of arc-annotated sequences [36], [35].

At the international level, a large number of groups develop similar researches on bioinformatics for noncoding RNAs, amongst them the Institut für theoretische Chemie, from Vienna University (I. Hofacker), the group of Practical Computer Science from Universität Bielefeld (R. Giegerich), the Bioinformatics Research Group from University of Copenhaguen (J. Gorodkin), the Computer Science Group from Ben Gurion University (M. Ziv-Ukelson). The specificity of SEQUOIA lies on a strong algorithmic background along with realistic biology models.

3.4. Non ribosomal peptides synthesis

This theme started in 2003 by informal discussions between members of the ProbioGEM laboratory (biological processes, enzymatic and microbial engineering) from Université Lille 1 and members of our team. Non ribosomal peptides (NRPs) are small molecules (2 to 50 residues) that have a branching or cyclic structure and that can incorporate many non-standard amino acids. We realized that few bioinformatics tools dedicated to (NRPs) exist and that filling this lack should be an exciting challenge. The work really started with the coming of Ségolène Caboche in spring 2006 for her master thesis, followed by a PhD thesis, defended in September 2009. The objective of the thesis was to design a computational resource for working with nonribosomal peptides, called NORINE.

NORINE is available from summer 2006 at the website http://bioinfo.lifl.fr/norine/. This first release has been announced in the Nucleic Acids Research 2008 database issue [1]. Today, the database contains 1116 nonribosomal peptides extracted from scientific literature with manually curated annotations such as biological activity, producing organisms or bibliographic references.

The database can be queried to search for peptides through their annotations as well as through their monomeric structure (structural pattern search or structure comparison). It also contains a section dedicated to the 506 different monomers incorporated into the peptides stored in NORINE. We developed and implemented in NORINE several efficient algorithms to compare NRP molecules represented as non-oriented labeled graphs. This led us to work on difficult computational problems in the field of graph isomorphism. We proposed an efficient algorithm that uses a variant of so-called compatibility graphs to search for all peptides containing a given structural pattern. The algorithm has been published in [14], also containing some real-life biological examples of using this features for mining NORINE for biologically useful information.

Currently, Norine is becoming the international reference resource for NRPs.

3.5. High-performance bioinformatics

Sequence analysis often make use of intensive computing. Examples include algorithms based on the dynamic programing paradigm, or algorithms on efficient data structures such suffix trees or suffix arrays. Performance comes with better algorithms, but also with better supports of execution using *parallelism*. This theme started with a collaboration with Symbiose team (INRIA Rennes) on sequence filtering methods with a reconfigurable architecture (ARC INRIA Flash 2006/2007) [55], [56]. Since 2008, we started to work on Graphics Processing Units (GPUs).

GPUs were used in bioinformatics since 2005 for phylogenetic studies [39], and for multiple sequence alignment based on an optimized Smith-Waterman implementation [48]. Recent papers provide speedups on bioinformatics applications involving suffix trees [60], Smith-Waterman comparisons [51], or RNA folding [58]. The best speed-ups are obtained when combining precise algorithmic analyses with a knowledge of the computing architectures. This is especially true with the memory hierarchy: the algorithms have to find a good balance between using large (and slow) global memories and some fast (but small) local memories.

Some other teams develop similar research on parallel bioinformatics, such as for example groups in Rennes (D. Lavenier), in the Nanyang Technological University, Singapore (B. Schmidt), in the University of Warsaw (W. Rudnicki), and in the Iowa State University (S. Aluru, J. Zola). This field of research is still in expansion.

4. Software

4.1. Introduction

We would like to stress the importance of software development in our team. All themes are equally concerned, and software dissemination is a fruitful ingredient of SEQUOIA. A special attention is paid to robustness, validation on biological data and availability. We maintain a web server accessible via http://bioinfo.lifl.fr/ for distributing our software and executing it through friendly web interfaces.

We give a list of tools that have been developed recently in the team, and that are in connection with our research program presented here. Many of them have been cited in the preceding paragraphs. These tools are likely to be enriched or improved within the next years.

4.2. YASS – local homology search

YASS is a software for computing similarity regions in genomic sequences, that is based on our work on novel seeding techniques. It is accompanied by HEDERA and IEDERA software programs that implement the algorithms of [45], [46] for seed optimization. The main features that distinguish YASS from its main competitors, such as PATTERNHUNTER, are the use of transition-constrained seeds, a powerful technique for designing those seeds, and an efficient implementation.

4.3. Carnac – RNA structure prediction

The CARNAC program implements our idea for RNA structure prediction presented in Section 3.3. It is written in C and is distributed under the Cecill license. The web interface offers 2D visualisation tools with Naview and alignment functionalities. It has proven to be very fast and very specific compared to its competitors.

4.4. Gardenia – RNA structure alignment

GARDENIA is a software for RNA structure alignment, based on our ideas presented in Section 3.3 on arc-annotated sequence models. It is enriched with an evolutionary model taking into account affine gap penalties, constraints coming from the primary structure, fast alignment for similar sequences, and local search. It also proposes options for pairwise alignment with tree edit distance, alignment of trees and multiple sequence alignment based on the ClustalW-like progressive alignment strategy. It comes with the RIBOSUM score matrices. It is written in C and is distributed under the GPL license.

4.5. Regliss – RNA locally optimal structures

REGLISS is a tool that studies the energy landscape of a given RNA sequence by considering locally optimal structures. Locally optimal structures are thermodynamically stable structures that are maximal for inclusion: they cannot be extended without producing a conflict between base pairs in the secondary structure, or increasing the free energy. The tool generates all locally optimal structures in a given sequence. Moreover, REGLISS can be used to explore the neighborhood of structures through an energy landscape graph.

4.6. Protea – coding sequence identification

PROTEA is a software for identifying evolutionary conserved coding sequences using a comparative analysis of genomic sequences. The rationale behind our method is that protein coding DNA sequences should feature mutations that are consistent with the genetic code and that tend to preserve the function of the translated amino acid sequence. The algorithm takes advantage of a specific substitution pattern of coding sequences together with the consistency of reading frames showing the best sequence similarity. This idea is original, and provides a complementary point of view to most of gene finders that are based on sequence composition bias or homology. The implementation uses graph-theoretical models to combine pairwise alignments and estimates the significancy of the conservation of the reading frames. This work appeared in [15]. PROTEA is distributed under the Cecill license.

4.7. Magnolia – multiple alignment via comparative analysis

MAGNOLIA is a new method to construct multiple alignments of nucleic acid sequences. It exploits our ideas coming from comparative analysis presented for noncoding RNAs and for coding sequences. Nucleic acids sequences are recognized to be hard to align because similarity is often reduced at the DNA level. The idea implemented is MAGNOLIA is to take into account the putative function of the sequences and to incorporate this functional information into the alignment. It takes as input a set of unaligned nucleic acids sequences, classifies the sequences either as coding RNAs using PROTEA or noncoding RNAs using CARNAC and produces a multiple sequence alignment based on the appropriate evolutionary pattern. When sequences are predicted as coding, then the multiple alignment relies on the putative amino-acid sequences with CLUSTALW. When sequences are predicted as noncoding, then the multiple alignment relies on the putative conserved secondary structure with GARDENIA. MAGNOLIA has been evaluated on large experimental data sets and appeared in [41].

4.8. CG-seq – comparative genomics

Genome annotation by comparative genomics is a common motivating background of several research subjects in the team: sequence processing, noncoding RNAs,...In this perspective, we have developed the CG-SEQ package, that gathers several tools for this task. The purpose of CG-seq is to identify functional regions in a genomic sequence by comparative analysis using multispecies comparison. It takes as input a genomic sequence and a set of other sequences coming from a variety of species to be compared against the target sequences, and proceeds in four steps. CG-SEQ includes YASS and CARNAC software, and is also designed to run with BLAST and RNAZ. By now, CG-seq is focused on noncoding RNAs. It could be extended to coding regions using PROTEA. CG-SEQ is available in a command-line version, and with a GUI written in Java. It is distributed under the GPL license.

4.9. RNAspace – a platform for noncoding RNA annotation

RNAspace is an open source platform born from a national collaborative initiative. Its goal is to develop and integrate functionalities allowing structural and functional noncoding RNA annotation (see Section 3.3).

RNAspace is written in Python with the template engine Cheetah and the object-oriented HTTP framework Cherrypy. It totalizes more than 17,000 lines of code. It is currently made available through a web site that has been opened to the scientific community for a few months: http://www.rnaspace.org, and it is distributed under the GPL licence. The project has been awarded by the national IBISA label in autumn 2009[1].

[1] IBISA is a French consortium for evaluating and funding national technological platforms in life sciences.

4.10. Norine – a resource for nonribosomal peptides

NORINE is a public computational resource for working with nonribosomal peptides. The database is managed with PostgreSQL and contains 21 tables. The web interface, including the tools for comparing NRPs, is developed in JSP (JavaServer Pages) and totalizes around 13,000 lines of code. It has been registered at the *Agence pour la Protection des Programmes*. NORINE is queried from all around the world. Our main users come from the United States of America, China and Germany where renowned biology laboratories work on nonribosomal peptides or on their synthetases.

5. New Results

5.1. Sequence processing and Next-Generation Sequencing

Participants: Mathieu Giraud, Marta Girdea, Gregory Kucherov, Laurent Noé.

5.1.1. Next Generation Sequencing

We have started last year a new collaboration with Institut Curie (Paris). We are working on a new method of genome mapping of reads issued from the SOLiD^{TM} sequencing technology. Although there is a number of available software programs for read mapping (ShRiMP, Bowtie, MAQ, PerM, ZOOM, ...), our first results show that we are able to obtain a better performance by using a finer modeling of read coding properties and read qualities. This is done by applying our seed technology, assisted by the IEDERA program (see Section 4.1). While this work is only in its very beginning, preliminary results have been presented in [29]. It should be developed in the PhD thesis of M. Gîrdea.

We also have continued our experience with parallel sequence filtering methods [55], pursuing this direction on NGS: We have recently acquired knowledge of SIMD techniques for fast parallel bandwith alignments that are fruitful on NGS problems.

5.1.2. Runs and palindromes

We continued algorithmic studies on palindromic and periodic structures in words (sequences). In 2009, we proposed a method to exactly count $\rho_p(n)$, the maximal number of runs with period at most p, and provided the first exact limits for some microruns [16].

5.2. Protein coding sequences

Participants: Arnaud Fontaine, Marta Girdea, Gregory Kucherov, Laurent Noé, Hélène Touzet.

5.2.1. Back-translation

Back-translation is the process of computing the putative DNA sequences that encode a given protein. Despite the fact that the number of back-translated sequences increase exponentialy with the size of the protein, such sequences are usefull, especially when dealing with frameshift proteins.

For the last two years, we have been interested by such approach to detect remote homologies. We have proposed an efficient dynamic programming alignment algorithm over the complete set of putative DNA sequences of each protein, to determine the two putative DNA sequences according to a scoring scheme designed to reflect the most probable evolutionary process. This allows us to uncover evolutionary information that is not captured by traditional alignment methods, which is confirmed by biologically significant examples.

The results have been published in the WABI workshop this year ([22]), and the extended version of this work is currenly accepted to the Algorithms in Molecular Biology journal. A web interface of the tool developped within this framework is now proposed at http://bioinfo.lifl.fr/path/

5.2.2. Computational identification of protein-coding sequences

Gene prediction is an essential step in understanding the genome of a species once it has been sequenced. For that, a promising direction in current research on gene finding is a comparative genomics approach. We designed a novel approach to identify evolutionary conserved protein-coding sequences in genomes. The rationale behind the method is that protein coding sequences should feature mutations that are consistent with the genetic code and that tend to preserve the function of the translated amino acid sequence. The algorithm takes advantage of the specific substitution pattern of coding sequences together with the consistency of reading frames. It has been implemented in a software called PROTEA. We have conducted a large scale analysis on thousands of conserved elements across eighteen eukaryotic genomes, including the Human genome. This experiment reveals the existence of new putative protein-coding sequences. Most of them are likely to be involved in alternative splicing transcripts, or to correspond to unannotated exons of predicted genes. This work appeared in [15].

5.3. Comparative analysis and Noncoding RNAs

Participants: Mathieu Giraud, Benjamin Grenier-Boley, Antoine de Monte, Azadeh Saffarian, Hélène Touzet.

5.3.1. Finding ncRNAs by comparative analysis

We have deviced a new method to find ncRNAs in genomes by comparative analysis. First, sequences are preprocessed for masking known annotated features, redundancy, ...Then the target sequence is compared to all other sequences to detect similar sequences across species. Pairwise alignments are combined into clusters of conserved regions. For that, the algorithm searches for regions whose conservation is supported by a significantly high number of pairwise alignments. Finally, conserved sequences are investigated by inspection of evolutionary patterns to identify conserved consensus secondary structures. This work has been implemented in the CG-seq software, and should to give rise to publication.

5.3.2. RNA pattern matching

Given a description for an RNA family, the goal is to identify all its potential occurrences on a genomic sequence, in a database or in a large set of small sequences. Stochastic context-free grammars turned out to be successful models for that, both in terms of sensitivity and specificity [52]. However, a high computational complexity of the related dynamic programming algorithms limits their practical application. More generally, an exhaustive benchmarking for RNA pattern matching shows that existing methods should compromise between efficiency and sensitivity, and even the fastest programs are not suitable for a genome-scale analysis [42]. We are currently working on filtering strategies, exploiting the approximate relative location of structural elements within the RNA motif, as well as conserved motifs within the alignment [64]. This filtering approach is intended to be used complementarily to exact methods as a preprocessing of the sequence. On a longer term, it also opens the way to the creation of new indexing structures whose goal is to store genomic data and to speed up RNA motif queries on this data.

5.3.3. RNA locally optimal structures

When the structure of a noncoding RNA is not known, it is still possible to enhance the search by considering the set of all plausible secondary structures. This gives rise to a new problem, that we call the multi-structure matching. This is the subject of the thesis of A. Saffarian. Her work aims at defining better data models for the set of all secondary structures of a given RNA, including suboptimal and locally optimal ones, and associated efficient algorithms for pattern matching [27].

5.3.4. The RNAspace open-source platform

Besides these theoretical issues, we are part of a new consortium for a national collaborative open-source platform devoted to noncoding RNA analysis, called RNASPACE. The project is conducted in collaboration with INRA Toulouse[2] (Christine Gaspin) and Institut de Génétique et Microbiologie de l'Université Paris

[2]INRA: French National Institute for Agricultural Research

Sud (Daniel Gautheret). Its goal is to develop and integrate functionalities allowing structural and functional noncoding RNA annotation. The platform allows the user to run a set of tools including most appropriate noncoding RNA gene finders, to integrate results and to explore and analyse RNA gene candidates. SEQUOIA is involved in RNAspace as a main contributor to this project. This is also a stepping stone for other tools developed in the team: CARNAC, GARDENIA, YASS and CG-SEQ are made available in the first release of RNAspace.

5.4. Non ribosomal peptides synthesis

Participants: Ségolène Caboche, Gregory Kucherov, Maude Pupin.

5.4.1. Study of NRPs monomeric composition

The rationale of a large-scale study of monomers of nonribosomal peptides is that such a study was not possible before the advent of NORINE, because it requires a complete list of these peptides annotated with the list of composing monomers. We discovered a significant difference in monomeric compositions of bacterial and fungi NRPs. Moreover, the monomeric composition can often be used as a signature of the producing organism or the biological activity of the NRP. A prediction software, based on these observations, has been created. Two papers describing these findings are in preparation.

5.4.2. Thesis

Ségolène Caboche obtain her PhD thesis in computational science the 8th of September [12].

5.4.3. External recognition

We have been contacted by the Worldwide Protein Data Bank [34] (wwPDB) that maintains a unique freely available archive of macromolecular structural data. wwPDB started to use NORINE entry identifiers as external references along with the UniProt [40] (Universal Protein Resource) accession codes for gene product sequences. Recently, NORINE has been described in an issue of Methods in Enzymology dedicated to microbial natural products biosynthesis [33].

5.5. High-performance bioinformatics

Participants: Mathieu Giraud, Tuan Tu Tran, Jean-Stéphane Varré.

5.5.1. Parallelisation of PWM algorithms

We studied some parallelisations on different Position-Weight-Matrices (PWMs) algorithms. The algorithms connecting the score threshold and the Pvalue are NP-complete and rely on the enumeration on a large set of scores or words [11]. While the basic algorithm for finding occurrences of a PWM uncovered the best speedup, we proposed a new algorithm for the score/Pvalue computation that is more suitable for parallelization. We realized a prototype with the CUDA libraries, and report for the different problems speedups of $21\times$ and $77\times$ on the GPU Nvidia GTX 280 [21].

5.5.2. GPU Parallelization of ADP

Algebraic Dynamic Programming (ADP) is a framework developed at U. Bielefeld (Germany) to encode a broad range of optimization problems, including common bioinformatics problems like RNA folding or pairwise sequence alignment. During the stay of P. Steffen in Lille, we proposed a first generic parallelization of this framework. Depending on the application, we report speedups ranging from $6.1\times$ to $25.8\times$ on a Nvidia GTX 280 [25].

5.5.3. Biomanycores

We started a project which intends to provide facilities for bioinformaticians to use parallel algorithms developed for the GPUs. The idea is to provide a repository for algorithms, notably written in OpenCL, and easy access through interfaces in the BioJava, BioPerl and Biopython frameworks. The project has been presented at the annual conference of bioinformatics open-source software (BOSC) [31].

5.6. Genome rearrangements

Participants: Aude Darracq, Jean-Stéphane Varré.

We ended the study of the rearrangement events of maize mitogenomes. Thanks to the analysis of the data we showed that tandem duplications with loss could be a common evolutionary mechanism in both plant and animal mitochondrial genomes. Such events are well-known in animals. We then proposed a method to deal with duplicated genes under this assumption and built for the first time an evolutionary scenario between mitochondrial genomes in higher plants (article under revision in BMC Genomics).

The sequencing project of beet mitochondrial genomes revealed a high difficulty in the assembling process because of a large number of duplicates in those genomes. The tandem duplication with loss hypothesis suggested us biological experiments (PCRs) allowing for contig assembly. The genomes will be available during November, 2009.

6. Contracts and Grants with Industry

6.1. NVIDIA

Since 2008, we are in contacts with NVIDIA, one of the leading companies in producing graphics processing units (GPUs). In 2008, NVIDIA gave to the team a Tesla S870 computing server (rack 1U with 4 GPUs) to test our parallel algorithms. In 2009, NVIDIA provided a Professor Partnership grant to the team.

7. Other Grants and Activities

7.1. Regional initiatives and cooperations

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

- The project on *nonribosomal peptide synthesis* is based on a collaboration with the ProBioGEM laboratory (*Laboratoire des Procédés Biologiques Génie Enzymatique et Microbien*), headed by Pr. Dhulster, University Lille 1. This laboratory develops methods to produce and extract active peptides in agriculture or food. The PhD work of Ségolène Caboche defended this year was co-supervised by Valérie Leclère from ProBioGem. A PhD work started on this subject in 2008: Aurélien Vanvlassenbroeck is working at ProBioGEM and is co-supervised by Maude Pupin.
- We collaborate with the *Laboratoire de Génétique et Évolution des Populations Végétales* (UMR CNRS 8016), Université de Lille 1 on the study of genomic rearrangements in the mitochondrial genome of higher plants. The goal is to identify evolutionary forces and molecular mechanisms that modeled the present diversity of mitochondrial genome at the species level, and in particular potentially active recombination sequences that have been used in the course of time. Data is acquired thanks to Genoscope projects (in beet and silene). A PhD student (Aude Darracq) is co-supervised on this subject.
- Our team is a member of the *PPF Bioinformatique*. This is an initiative of the University Lille 1 that coordinates public bioinformatics activities at the local level for the period 2006-09.

7.2. National initiatives and cooperations

7.2.1. National initiatives

We participate in the following national projects:

- ANR BRASERO (2007-2009). The project aims at providing relevant and efficient tools for the RNA comparison problem. The project is coordinated by A. Denise (now in AMIB INRIA team). Other leading partners are SEQUOIA, the bioinformatics group of LaBRI, Bordeaux (with P. Ferraro), and the BAMBOO project of INRIA Rhône-Alpes (with M.-F. Sagot). The project also involves researchers from *Institut de Génétique et Microbiologie* (Orsay), *Centre de Génétique Moléculaire* (Gif-sur-Yvette) and *Maturation des ARN et Enzymologie Moléculaire* (Nancy).
- ANR CoCoGen (2008-2011). The goal of this project is to study new methods for comparison of complete genomes. The project is coordinated by E. Rivals (LIRMM, Montpellier). Others participants are MIG and UBLO teams of INRA (Jouy-en-Josas), INA-PG (Paris). The budget of this project is managed by the Montpellier partner. It covers travel fees to attend meetings.
- NCRNA, RNG-Renabi, national network for bioinformatics (2007-2009). The objective is to develop an open-source annotation platform for noncoding RNA genes (see RNAspace in Section 3.3). This project involves the bioinformatics platforms of Génopole Toulouse-Midi-Pyrénées and SEQUOIA, and is supervised by C. Gaspin (Toulouse-Midi-Pyrénées). New support is planned for 2010.
- working groups *Sequence analysis* and *Structural bioinformatics* of the multidisciplinary *GDR Molecular bioinformatics*[3].
- working group *Combinatoire des mots, algorithmique du texte et du génome* of the *GDR Informatique Mathématique*[4].

7.2.2. National cooperations

- Work on mapping SOLiD reads: Group of E. Barillot (Institut Curie, Paris)
- The following scientists were invited in the past year to give a talk at the team seminar: David Hot (Institut Pasteur de Lille), Guillaume Rizk (INRIA Rennes), Mikaël Salson (Université de Rouen), Peter Steffen (Universität Bielefeld), Martin Figeac (INSERM, Université Lille 2), Sylvain Sené (Université Grenoble 1), Christophe Pinchon (Institut Pasteur), Pavlos Antoniou (King's College London), Aïda Ouangraoua (Simon Fraser University), Samuel Blanquart (LIRMM), Géraldine Jean (LaBRI), Mathieu Raffinot (LIAFA)
- We collaborate for several years with the INRIA team SYMBIOSE (Rennes), with D. Lavenier and P. Peterlongo, on indexation and seed-based heuristics.

7.3. International initiatives and cooperations

- We are in contact with the group of Prof. Pavel Pevzner (Center for Algorithmic and Systems Biology at the University of California, San Diego), as they use NORINE for mass spectrometry data interpretation.
- We want to enlarge the biological area of NORINE to synthetases, the enzymes that synthetise the nonribosomal peptides. This work should benefit from a new international collaboration with Dr. Hranueli's lab[5], one of the main international contributor in the field of bioinformatics for synthetases [61]. We spent one week in his lab in April 2009, and they come one week in our lab in September 2009.

[3] http://www.gdr-bim.u-psud.fr
[4] http://www.gdr-im.fr/
[5] Faculty of Food Technology and Biotechnology, University of Zagreb

- Peter Steffen, from University Bielefeld (Germany), visited our group for one month in March 2009, and gave a talk in the LIFL seminar. He collaborates with M. Giraud on a GPU implementation for the ADP (algebraic dynamic programming) methodology. We submitted a proposal of a PROCOPE bilateral cooperation project for 2010 – 2011.
- Matthias Bernt, from University of Liepzig (Germany), visited our group for two weeks in May an June 2009. He collaborated with A. Darracq and J.-S. Varré on the expression power of common intervals in the process of building phylogenies considering genomic rearrangements.

8. Dissemination

8.1. Organization of workshops and seminars

8.1.1. CPM 2009

In summer 2009, the team hosted in Lille the 20th edition of the **CPM (Combinatorial Pattern Matching)** conference[6]. This established international conference gathered around 80 participants coming from all over the world.

8.1.2. Next Generation Sequencing

We organized a joint scientific meeting with Institut Pasteur de Lille on NGS technologies in March 2009 (70 participants).

8.1.3. GTGC working group

J.-S. Varré is one of the committee members of the national GTGC working group[7] (Comparative Genomics Working Group) created in 2005. The group organizes one or two seminar sessions per year on comparative genomics.

8.1.4. Arena working group

H.Touzet is one of the committee members of the national ARENA working group (bioinformatics of noncoding RNAs).

8.1.5. INRIA Lille GPGPU working group

M. Giraud organizes since September 2008 a working group on “general purpose computing on GPUs”, with 2-3 meetings a month. This working group gathers 10 scientists from four different EPI in INRIA Lille Nord Europe.

8.1.6. Journées au vert

On May, 4-5, 2009, we organized a team two-days seminar in Saint Valery sur Somme in order to discuss current and future research projects carried out in the group.

8.2. Editorial and reviewing activities

- Editorial Board of BMC Algorithms for Molecular Biology (G. Kucherov)
- Program committee of PSI 2009 (G. Kucherov), CPM 2009 (G. Kucherov, H. Touzet), JOBIM 2009 (G. Kucherov, L. Noé), ICCS/WEPA 2010 (M. Giraud) JOBIM 2010 (H. Touzet)

[6] http://bioinfo.lifl.fr/cpm09
[7] http://www.lina.univ-nantes.fr/conf/gtgc2009/GTGC.html

- Reviewer for the journals Journal of Computer and System Sciences (G. Kucherov), Int. Journal of Foundations of Computer Science (M. Giraud), Nucleic Acids Research (H. Touzet, S. Caboche, J.-S. Varré), Theoretical Computer Science (G. Kucherov), IEEE Transactions on Bioinformatics and Computational Biology (H. Touzet) Biofutur (H. Touzet) Food Technology and Biotechnology (M. Pupin)
- Reviewer for the conferences SODA 2009 (G. Kucherov), STACS 2009 (M. Giraud, G. Kucherov). PSC 2009 (M. Giraud), IWOCA 2009 (M. Giraud), FCT 2009 (H. Touzet)

8.3. Miscellaneous activities

- Jury of the PhD theses of Isabelle da Piedade, University of Copenhaguen (H. Touzet), Ana Kozomara, INRA Toulouse (H. Touzet), Alexandru-Adrian Tantar, Université Lille 1 (H. Touzet)
- Reviewers for the french ministry program ANR (J.-S. Varré, H. Touzet)
- Reviewers for the INRIA "Équipes Associées" program (G. Kucherov, J.-S. Varré)
- Reviewer for the INRIA Commission d'Évaluation (H. Touzet)
- Reviewer for the Israel Science Foundation (H. Touzet)

8.4. Meetings attended and talks

8.4.1. International Conferences

- LATA 2009, *Language and Automata Theory and Applications*, Tarragona, Spain, April 2009 (J.-S. Varré [24])
- ISPDC 2009, *Int. Symposium on Parallel and Distributed Computing*, Lisboa, Portugal, July 2009 (M. Giraud [21])
- PPAM/PBC 2009, *Parallel Processing and Applied Mathematics, Parallel BioComputing Workshop*, Wroclaw, Poland, September 2009 (M. Giraud [25])
- WABI 2009, *Workshop on Algorithms in Bioinformatics*, Philadelphia, US, September 2009 (L. Noé [22], [29])

8.4.2. National Conferences

- JOBIM 2009, *Journées Ouvertes Biologie Mathématique Informatique Biologie*, Nantes, June 2009 (A .Saffarian [27], A. Darracq [26], M. Pupin)

8.4.3. Talks, meetings, seminars

- BOSC 2009, *Bioinformatics Open Source Conference*, Stockholm, June 2009 (J.-S. Varré [31])

8.5. Teaching activities

Our research work finds also its expression in a strong commitment in pedagogical activities at the University Lille 1. For several years, members of the project have been playing a leading role in the development and the promotion of bioinformatics (more than 400 teaching hours per year). We are involved in several graduate diplomas (research master degree) in computer science and biology (*master protéomique, master biologie-santé, master génie cellulaire et moléculaire, master interface physique-chimie*) in an Engineering School (Polytech'Lille), as well as in permanent education (for researchers, engineers and technicians).

8.5.1. Lectures on bioinformatics, University of Lille 1

- Organization of a lecture series on *Algorithms and computational biology*, master in computer science (M2), 15h (L. Noé, M. Pupin, H. Touzet)
- *Bioinformatics*, master génomique et protéomique (M1), 64h (L. Noé, M. Pupin, S. Caboche)
- *Bioinformatics*, master génomique et microbiologie (M1), 24h (M. Giraud)
- *Bioinformatics*, master protéomique (M2), 30h (M. Pupin)
- *Bioinformatics*, master génie cellulaire et moléculaire (M2), 30h (M. Pupin)
- *Bioinformatics*, master biologie-santé (M2), 14h (M. Pupin, A. Darracq)
- *Bioinformatics*, master from Polytech'Lille, 24h (M. Pupin, A. Darracq)

8.5.2. Teaching in computer science, University of Lille 1

- *Algorithmics*, second year IUT students, 40h (A. Fontaine)
- *Computers architecture*, first year IUT students, 24h (A. Fontaine)
- *Programming (Pascal)*, first year of bachelor, 36h (M. Pupin, L. Noé)
- *Programming (Ocaml, Prolog)*, third year of bachelor, 48h (L. Noé)
- *Networks*, third year of bachelor, 36h (L. Noé)
- *Operating systems architecture*, first year of master, 42h (L. Noé)
- *Professional project*, first year of master, 16h (M. Pupin)
- *Web technologies*, PhD students, 18h (M. Pupin)
- *Algorithmics*, second year of bachelor, 30h (A. Saffarian)

8.5.3. Other teaching duties

- *Graph theory*, second year of engineering school, 32h (A. Saffarian)

8.6. Administrative activities

- Member of the executive commitee of *GDR Molecular bioinformatics* (H. Touzet)
- Member of the GTAI INRIA committee (H. Touzet)
- Member of the INRIA evaluation commitee (M. Giraud)
- Member of hiring committee *(jury d'audition)* of INRIA Bordeaux and Grenoble (M. Giraud)
- Members of hiring committee *(Commission des Spécialistes)* of the University Lille 1 (H. Touzet and M. Pupin)
- Member of the Men/Women egality working group, University Lille 1 (M. Pupin)
- Head of PPF bioinformatics – University Lille 1 (H. Touzet)
- Member of the LIFL Laboratory council (H. Touzet)
- Member of the INRIA Lille center commitee (J.-S. Varré)

9. Bibliography

Major publications by the team in recent years

[1] S. CABOCHE, M. PUPIN, V. LECLÈRE, A. FONTAINE, P. JACQUES, G. KUCHEROV. *NORINE: a database of nonribosomal peptides*, in "Nucleic Acids Research", vol. Database Issue, Vol. 36, 2008, p. D326–331, http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D326.

[2] M. DEFRANCE, H. TOUZET. *Predicting transcription factor binding sites using local over-representation and comparative genomics*, in "BMC Bioinformatics", 2006, http://www.biomedcentral.com/1471-2105/7/396/abstract.

[3] G. DIDIER, I. LAPREVOTTE, M. PUPIN, A. HENAUT. *Local decoding of sequences and alignment-free comparison.*, in "Journal of Computational Biology", vol. 13, n° 8, 2006, p. 1465–1476, http://dx.doi.org/10.1089/cmb.2006.13.1465.

[4] S. DULUCQ, H. TOUZET. *Decomposition algorithms for the tree edit distance problem*, in "Journal of Discrete Algorithms", 2005, p. 448-471, http://dx.doi.org/10.1016/j.jda.2004.08.018.

[5] M. FIGEAC, J.-S. VARRÉ. *Sorting By Reversals with Common Intervals*, in "Proceedings of the 4th International Workshop Algorithms in Bioinformatics (WABI 2004), Bergen, Norway, September 17-21, 2004", Lecture Notes in Computer Sciences, vol. 3240, Springer Verlag, 2004, p. 26-37.

[6] R. KOLPAKOV, G. KUCHEROV. *Identification of periodic structures in words*, in "Applied combinatorics on words", J. BERSTEL, D. PERRIN (editors), Lothaire books, vol. Encyclopedia of Mathematics and its Applications, vol. 104, chap. 8, Cambridge University Press, 2005, p. 430–477, http://www-igm.univ-mlv.fr/~berstel/Lothaire/index.html.

[7] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *Multi-seed lossless filtration*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", vol. 2, n° 1, January-March 2005, p. 51–61.

[8] D. LAVENIER, M. GIRAUD. *Bioinformatics Applications*, in "Reconfigurable Computing: Accelerating Computation with Field-Programmable Gate Arrays", Springer, 2005, http://dx.doi.org/10.1007/0-387-26106-0_8.

[9] A. LIEFOOGHE, H. TOUZET, J.-S. VARRÉ. *Large Scale Matching for Position Weight Matrices.*, in "Proceedings 17th Annual Symposium on Combinatorial Pattern Matching (CPM)", Lecture Notes in Computer Science, vol. 4009, Springer Verlag, 2006, p. 401–412, http://www.springerlink.com/content/7113757vj6205067/.

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

[11] H. TOUZET, J.-S. VARRÉ. *Efficient and accurate P-value computation for Position Weight Matrices*, in "Algorithms for Molecular Biology", vol. 2, n° 15, 2007.

Year Publications

Doctoral Dissertations and Habilitation Theses

[12] S. CABOCHE. *Mise en place d'une plateforme logicielle pour l'analyse des peptides non-ribosomiaux*, Université de Lille 1, 2009, Ph. D. Thesis.

[13] A. FONTAINE. *Classification d'ARN codants et d'ARN non-codants*, Université de Lille 1, 2009, http://www2.lifl.fr/~fontaina/These.pdf, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

[14] S. CABOCHE, M. PUPIN, V. LECLÈRE, P. JACQUES, G. KUCHEROV. *Structural pattern matching of nonribosomal peptides*, in "BMC Structural Biology", vol. 9:15, March 18 2009, http://www.biomedcentral.com/1472-6807/9/15.

[15] A. FONTAINE, H. TOUZET. *Computational identification of protein-coding sequences by comparative analysis*, in "International Journal of Data Mining and Bioinformatics", vol. 3, n° 2, 2009, p. 160–176, http://www.inderscience.com/search/index.php?action=record&rec_id=24849.

[16] M. GIRAUD. *Asymptotic behavior of the number of runs and microruns*, in "Information and Computation", vol. 207, n° 11, 2009, p. 1221-1228, http://dx.doi.org/10.1016/j.ic.2009.02.007.

[17] R. KOLPAKOV, G. KUCHEROV. *Searching for gapped palindromes*, in "Theoretical Computer Science", vol. 410, n° 51, 2009, p. 5365–5373RU.

[18] A. OUANGRAOUA, P. FERRARO. *A constrained edit distance algorithm between semi-ordered trees*, in "Theor. Comput. Sci.", vol. 410, n° 8-10, 2009, p. 837-846.

[19] A. OUANGRAOUA, P. FERRARO. *A new constrained edit distance between quotiented ordered trees*, in "J. Discrete Algorithms", vol. 7, n° 1, 2009, p. 78-89.

[20] M. ROYTBERG, A. GAMBIN, L. NOÉ, S. LASOTA, E. FURLETOVA, E. SZCZUREK, G. KUCHEROV. *On subset seeds for protein alignment*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", vol. 6, n° 3, 2009, p. 483–494, http://www.lifl.fr/~noe/files/pp_TCBB09_preprint.pdfRUPL.

International Peer-Reviewed Conference/Proceedings

[21] M. GIRAUD, J.-S. VARRÉ. *Parallel Position Weight Matrices Algorithms*, in "Proceedings of the 8th International Symposium on Parallel and Distributed Computing, ISPDC'09", 2009, p. 65-69, http://dx.doi.org/10.1109/ISPDC.2009.31.

[22] M. GÎRDEA, G. KUCHEROV, L. NOÉ. *Back-translation for discovering distant protein homologies*, in "Proceedings of the 9th International Workshop in Algorithms in Bioinformatics (WABI), Philadelphia (USA)", S. SALZBERG, T. WARNOW (editors), Lecture Notes in Computer Science, vol. 5724, Springer Verlag, September 2009, p. 108–120, http://www.springerlink.com/content/3236004m84465n7j/.

[23] E. HARRIS, T. LECROQ, G. KUCHEROV, S. LONARDI. *CPM's 20th Anniversary: A Statistical Retrospective*, in "Proceedings of the 20th Annual Combinatorial Pattern Matching Symposium (CPM), Lille (France), June 22-24, 2009", Lecture Notes in Computer Science, vol. 5577, Springer Verlag, 2009, p. 1-11ITUS.

[24] A. LIEFOOGHE, H. TOUZET, J.-S. VARRÉ. *Self-overlapping occurrences and Knuth-Morris-Pratt algorithm for weighted matching*, in "Proceedings of the 3rd International Conference on Language and Automata Theory and Applications, April 2-8, 2009 - Tarragona, Spain", vol. 5457, 2009, p. 481–492, http://www.springerlink.com/content/p6q458l886880v45/.

[25] P. STEFFEN, R. GIEGERICH, M. GIRAUD. *GPU Parallelization of Algebraic Dynamic Programming*, in "Parallel Processing and Applied Mathematics / Parallel Biocomputing Conference (PPAM / PBC 09)", 2009, to appear DE .

National Peer-Reviewed Conference/Proceedings

[26] A. DARRACQ, J.-S. VARRÉ, P. TOUZET. *A Study of Genomic Rearrangements in Maize Mitochondrial Genomes*, in "Journées Ouvertes Biologie Informatique Mathématiques (JOBIM 2009)", 2009, poster.

[27] A. SAFFARIAN, M. GIRAUD, H. TOUZET. *Paysage d'énergie et structures localement optimales d'un ARN*, in "Journées Ouvertes Biologie Informatique Mathématiques (JOBIM 2009) (poster)", 2009, poster.

Workshops without Proceedings

[28] A. DARRACQ, J.-S. VARRÉ, P. TOUZET. *A Study of Genomic Rearrangements in Maize Mitochondrial Genomes*, in "The 17th annual meeting of the Society for Molecular Biology and Evolution (SMBE 2009)", 2009, poster.

[29] M. GÎRDEA, L. NOÉ, G. KUCHEROV. *Read Mapping Tool for AB SOLiD Data*, in "9th Workshop on Algorithms in Bioinformatics (WABI), Philadelphia (USA), September 12-13, 2009", September 2009, http://www.lifl.fr/~noe/files/poster_WABI09.abstract.pdf, poster.

[30] R. URICARU, C. MICHOTEY, L. NOÉ, H. CHIAPELLO, É. RIVALS. *Improved sensitivity and reliability of anchor based genome alignment*, in "Proceedings of the 10th Open Days in Biology, Computer Science and Mathematics (JOBIM), June 9-11, 2009, Nantes (France)", 2009, p. 31–36, http://www.lirmm.fr/~uricaru/articles/jobim26.pdf.

[31] J.-S. VARRÉ, S. JANOT, M. GIRAUD. *Biomanycores, a repository of interoperable open-source code for many-cores bioinformatics*, in "Bioinformatics Open Source Conference (BOSC)", 2009.

Books or Proceedings Editing

[32] G. KUCHEROV, E. UKKONEN (editors). *Combinatorial Pattern Matching: 20th Annual Symposium (CPM), Lille (France), June 2009. Proceedings*, Lecture Notes in Computer Science, vol. 5577, Springer, 2009RUFI.

References in notes

[33] B. O. BACHMANN, J. RAVEL. *Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data*, in "Methods in Enzymology", vol. 458, 2009, p. 181–217.

[34] H. BERMAN, K. HENRICK, H. NAKAMURA, J. L. MARKLEY. *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data*, in "Nucleic Acids Research", vol. 35, n^{o} Database issue, 2007, p. D301–303.

[35] G. BLIN, A. DENISE, S. DULUCQ, C. HERRBACH, H. TOUZET. *Alignment of RNA structures*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", 2008.

[36] G. BLIN, H. TOUZET. *How to Compare Arc-Annotated Sequences: The Alignment Hierarchy*, in "13th International Symposium on String Processing and Information Retrieval (SPIRE)", Lecture Notes in Computer Science, vol. 4209, Springer Verlag, 2006, p. 291–303, http://www.springerlink.com/content/4k37q116j2720832/.

[37] D. G. BROWN. *A survey of seeding for sequence alignments*, in "Bioinformatics Algorithms: Techniques and Applications", I. MANDOIU, A. ZELIKOVSKY (editors), J. Wiley and Sons, 2008.

[38] S. BURKHARDT, J. KÄRKKÄINEN. *Better Filtering with Gapped q-Grams*, in "Proceedings of CPM'01", LNCS, vol. 2089, Springer-Verlag, 2001, p. 73–85, http://www.springerlink.com/index/GYKW51MPJQNWRMQX.pdf.

[39] M. CHARALAMBOUS, P. TRANCOSO, A. STAMATAKIS. *Initial Experiences Porting a Bioinformatics Application to a Graphics Processor*, in "Adv. in Informatics", 2005, p. 415–425.

[40] UNIPROT. CONSORTIUM. *The Universal Protein Resource (UniProt) 2009*, in "Nucleic Acids Research", vol. 37, n^{o} Database issue, 2009, p. D169–74.

[41] A. FONTAINE, A. DE MONTE, H. TOUZET. *MAGNOLIA: multiple alignment of protein-coding and structural RNA sequences*, in "Nucleic Acids Research", vol. Web Server Issue, Vol 36, n^{o} suppl 2, 2008, p. W14-W18, http://nar.oxfordjournals.org/cgi/content/full/gkn321.

[42] E. FREYHULT, J. BOLLBACK, P. GARDNER. *Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA*, in "Genome Research", vol. 17, 2007, p. 117 – 125.

[43] P. GARDNER, R. GIEGERICH. *A comprehensive comparison of comparative RNA structure prediction approaches*, in "BMC Bioinformatics", vol. 5(140), 2004.

[44] P. GELSINGER. *Microprocessors for the New Millennium: Challenges, Opportunities, and New Frontiers*, in "IEEE International Solid State Circuits Conference (ISSCC 2001)", 2001, p. 22-25.

[45] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *A unifying framework for seed sensitivity and its application to subset seeds*, in "Journal of Bioinformatics and Computational Biology", vol. 4, n^{o} 2, 2006, p. 553–569, http://www.worldscinet.com/jbcb/04/0402/S0219720006001977.html.

[46] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *Subset Seed Automaton*, in "12th International Conference on Implementation and Application of Automata (CIAA 07)", Lecture Notes in Computer Science, vol. 4783, Springer Verlag, 2007, p. 180–191, http://www.springerlink.com/content/y824l20554002756/.

[47] F. LIPMANN, W. GEVERS, H. KLEINKAUF, R. J. ROSKOSKI. *Polypeptide synthesis on protein templates: the enzymatic synthesis of gramicidin S and tyrocidine.*, in "Adv Enzymol Relat Areas Mol Biol", vol. 35, 1971, p. 1–34.

[48] W. LIU, B. SCHMIDT, G. VOSS, W. MÜLLER-WITTIG. *GPU-ClustalW: Using Graphics Hardware to Accelerate Multiple Sequence Alignment*, in "High Performance Computing (HiPC 2006), LNCS 4297", 2006, p. 363-374.

[49] B. MA, J. TROMP, M. LI. *PatternHunter: Faster and more sensitive homology search*, in "Bioinformatics", vol. 18, n^{o} 3, 2002, p. 440–445, http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/3/440.

[50] A. MACHADO-LIMA, H. PORTILLO, A. DURHAM. *Computational methods in noncoding RNA research*, in "Journal of Mathematical Biology", vol. 56 (1-2), 2008, p. 15-49.

[51] S. A. MANAVSKI, G. VALLE. *CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment*, in "BMC Bioinformatics", vol. 9 (S2), 2008, S10.

[52] E. S. NAWROCKI EP. *Query-dependent banding (QDB) for faster RNA similarity searches*, in "PLoS Comput Biology", vol. 3(3):e56, 2007.

[53] L. NOÉ, G. KUCHEROV. *YASS: Similarity Search in DNA Sequences*, n^{o} RR-4852, INRIA, 2003, Technical report.

[54] A. OUANGRAOUA, P. FERRARO, L. TICHIT, S. DULUCQ. *Local similarity between quotiented ordered trees*, in "J. Discrete Algorithms", vol. 5, n^{o} 1, 2007, p. 23-35.

[55] P. PETERLONGO, L. NOÉ, D. LAVENIER, G. LES GEORGES, J. JACQUES, G. KUCHEROV, M. GIRAUD. *Protein similarity search with subset seeds on a dedicated reconfigur able hardware*, in "Parallel Processing and Applied Mathematics / Parallel Biocomputi ng Conference (PPAM / PBC 07)", R. WYRZYKOWSKI, J. DONGARRA, K. KARCZEWSKI, J. WASNIEWSKI (editors), Lecture Notes in Computer Science (LNCS), vol. 4967, 2008, p. 1240-1248, http://www.lifl.fr/~giraud/publis/peterlongo-pbc-07.pdf.

[56] P. PETERLONGO, L. NOÉ, D. LAVENIER, V. H. NGUYEN, G. KUCHEROV, M. GIRAUD. *Optimal neighborhood indexing for protein similarity search*, in "BMC Bioinformatics", vol. 9, n^{o} 534, 2008, http://www.biomedcentral.com/1471-2105/9/534.

[57] M. POP, S. SALZBERG. *Bioinformatics challenges of new sequencing technology*, in "Trends Genetics", vol. 24(3), 2008, p. 142-9.

[58] G. RIZK, D. LAVENIER. *GPU accelerated RNA folding algorithm*, in "Using Emerging Parallel Architectures for Computational Science / International Conference on Computational Science (ICCS 2009)", 2009.

[59] M. ROYTBERG, A. GAMBIN, L. NOÉ, S. LASOTA, E. FURLETOVA, E. SZCZUREK, G. KUCHEROV. *Efficient seeding techniques for protein similarity search*, in "Proceedings of the 2nd Workshop on Algorithms in Molecular Biology (ALBIO'08), Vienna (Austria), July 7-9, 2008", M. ELLOUMI, J. KÜNG, M. LINIAL, R. MURPHY, K. SCHNEIDER, C. TOMA (editors), Communications in Computer and Information Science, vol. 13, Springer Verlag, 2008, p. 466–478, http://www.springerlink.com/content/m3560l36r573xjr5/.

[60] M. SCHATZ, C. TRAPNELL, A. DELCHER, A. VARSHNEY. *High-throughput sequence alignment using Graphics Processing Units.*, in "BMC Bioinformatics", vol. 8, 2007, 474.

[61] A. STARCEVIC, J. ZUCKO, J. SIMUNKOVIC, P. F. LONG, J. CULLUM, D. HRANUELI. *ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures*, in "Nucleic Acids Res.", vol. 36, 2008, p. 6882–6892.

[62] H. TOUZET. *Comparative analysis of RNA genes: the CaRNAc software*, vol. Methods in Molecular Biology, Special issue on comparative genomic s I, Humana Press, 2007, p. 465-473, http://hal.archives-ouvertes.fr/docs/00/17/85/57/PDF/comparativerna.pdf.

[63] H. TOUZET. *Comparing similar ordered trees in linear-time*, in "Journal of Discrete Algorithms", vol. 5, n^{o} 4, 2007, p. 696-705, http://linkinghub.elsevier.com/retrieve/pii/S1570866706000700.

[64] H. TOUZET. *Looking for RNA motifs, fast and well*, in "ARENA – non-coding RNA bioinformatics (Toulouse)", 2008, (talk).