INRIA

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 BIO

Table of contents

SEQUOIA *is a joint project-team with LIFL (CNRS-UMR 8022 and USTL/Lille 1 University).*

1. Team

Research Scientist

Hélène Touzet [CR CNRS, on maternity leave from July until November, Team leader from November 6, HdR]
Gregory Kucherov [DR CNRS, Team leader until November 6, on leave for Poncelet Institute, Moscow, HdR]
Jesper Jansson [CR INRIA, until April 2008]
Mathieu Giraud [CR CNRS]

Faculty Member

Jean-Stéphane Varré [MC, Université Lille 1, on leave at INRIA from September 2008]
Laurent Noé [MC, Université Lille 1]
Maude Pupin [MC, Université Lille 1]
Sylvain Guillemot [ATER, Université Lille 1, until April 2008]

Technical Staff

Antoine de Monte [Ingénieur, Université Lille 1, from July 2007]
Benjamin Grenier-Boley [Ingénieur Associé, INRIA, from September 2007]

PhD Student

Ségolène Caboche [INRIA/Region fellowship, from October 2006]
Aude Darracq [MESR fellowship, from October 2007]
Marta Girdea [INRIA CORDI fellowship, from October 2007]
Arnaud Fontaine [MESR fellowship, from October 2005]
Aude Liefooghe [MESR fellowship until August 2007, ATER Université Lille 1 from September 2007, defended in July 2008]
Azadeh Saffarian [MESR fellowship, from November 2007]

Post-Doctoral Fellow

Alban Mancheron [PostDoc, INRIA, until August 2008]

Visiting Scientist

Tetsuo Shibuya [University of Tokyo, February 14-27]
Peter Steffen [University of Bielefeld, October 8-11]

Administrative Assistant

Sandrine Catillon [INRIA]

2. Overall Objectives

2.1. Introduction

Keywords: *algorithmics*, *bioinformatics*, *comparative genomics*, *computational biology*, *discrete algorithms*, *genomic sequences*, *high-performance computing*, *non-coding RNAs*, *parallelisation*, *phylogenetics*, *protein sequences*, *regulation*, *sequence alignment*, *sequence analysis*, *word combinatorics*, *word statistics*.

For the last fifteen years bioinformatics has undergone a remarkable evolution and became a rich and very active research field. This advancement is associated with a breakthrough development of sequencing technologies that resulted in the availability of a large body of genomic data, as well as with the emergence of new high-throughput genomic, transcriptomic and proteomic technologies (DNA chips for monitoring gene expression, mass spectrometry, ...). Moreover, recent discoveries in molecular biology, such as a new understanding of the role of non-coding DNA, gave rise to new challenging bioinformatics problems. While modern bioinformatics features various mathematical models and methods, sequence analysis still remains one of its central components, especially with the huge amount of data produced by next-generation sequencers.

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.

2.2. Highlights of the year

- In July, Aude Liefooghe defended her PhD on an optimized search of regulatory patterns. In December, J.-S. Varré defended his Habilitation thesis.
- The team co-organized the French national bioinformatics conference JOBIM 2008. This multidisciplinary meeting gathered 350 researchers coming from computer science, biology, physics and mathematics.
- In 2008, Jesper Jansson left our team to go to the Ochanomizu University (Tokyo, Japan) to be with his wife. We thank Jesper for the 7 months he was in the team, and wish him the best for his scientific carrier as well as for his family.

3. Scientific Foundations

3.1. Comparative genomics

Comparative genomics is a paradigm that emerged from mass genome sequencing as well as from the appearance of bulks of other biological data. The rationale behind this paradigm is that deciphering certain biological mechanisms of genome expression can be made possible (or at least, drastically more efficient) by *comparing* genomic (or other) data of different organisms, rather than analyzing an individual organism. Besides revealing features common to different species and therefore likely to have a biological function, this approach can also take into account *evolutionary information* which is essential in modern bioinformatics studies.

To be put into practice, comparative genomics needs new computational tools. Those tools have to be not just simple improvements of existing ones but should be *qualitatively* more efficient in order to follow the exponential grow of available data. Most of the research subjects presented below follow this direction, i.e. aim at providing most efficient software tools for the large-scale genomic analysis.

3.2. Sequence similarity and repetitions

Keywords: *homology*, *repeat*, *sequence alignment*, *sequence similarity*.

A basic highly recurrent operation in manipulating biological sequences is comparing them in order to detect *similarity regions*. Being able to compute both quickly and precisely similar fragments of two sequences, or in a sequence and a database, is crucial for virtually all projects that deal with sequence data, and the corresponding software, such as the well-known BLAST package [40], is by far the most widely used bioinformatics software. Since the similarity search is the most low-level operation in sequence analysis, its efficiency is important for every upper level of analysis. An underlying idea common to these computations is that the presence of similar (*conserved*) sequences provides an evidence that these sequences bear a biological function; moreover, similar sequences are likely to correspond to similar biological functions and/or to a common evolutionary ancestor.

3.2.1. Efficient methods of sequence comparison

Several years ago, similarity search algorithms became subject of a remarkable improvement due to the invention of the concept of *spaced seeds*, first proposed in the context of DNA similarity search by the PATTERNHUNTER software [63]. The idea of spaced seeds results in a considerable gain in sensitivity of search, without loss of selectivity.

The advent of spaced seeds opened up a new research area as it raised a number of new questions: how to estimate the quality of spaced seeds? how to design them? how to define the class of possible seeds for a given comparison setting? how to efficiently implement them? etc. A number of papers have been devoted to these questions during the last years, see [44], [64], [83], [58], [48], [82], [55] to cite a few recent ones. We have been working in this area for several years and made several contributions of which the main one is the YASS software for DNA sequence alignment [69] [9] developed by group members (see Section 4.3).

To consider another aspect of this development, a spaced seed – or a set of spaced seeds – specifies a way of indexing a genomic sequence. This indexing scheme is more powerful than the one based on indexing contiguous words (k-mers or q-grams), as keys occurring at consecutive positions are more independent and therefore more information can possibly be drawn from the whole index without increasing its cost. On the other hand, reconfigurable computer architecture of type FPGA (see Section 3.6.3) provides possibilities for reducing the cost of accessing and manipulating sequence keys specified by spaced seeds.

Many other interesting issues arise in relation to spaced seeds and lead to various research problems. Without being exhaustive, let us mention the issue of statistical properties of keys in genomic sequences. A knowledge about those properties can help in designing efficient seeds. Another issue that is within our scope of interest is the design of *lossless seeds* i.e. seeds presenting 100% sensitivity. In contrast to the "usual" similarity search, where missing a certain (although small) number of interesting similarities is always admitted, some applications require *all* similarities to be found. The design of such seeds leads to difficult combinatorial questions that have recently been subject of several studies [6], [47], [68].

3.2.2. Repeated sequences in genomes

Sequences conserved within one sequence (e.g. one genome) are called *repeats*. It is well-known now that genomic sequences are highly repeated: for example, about a half of the human genome is composed of repeated occurrences of some significant-length sequences. Those sequences have very different syntactic characteristics (such as length or relative occurrence of repeated copies) and different (often unknown) biological functions. Moreover, *tandem repeats* have a particular consecutive structure that reflects yet different biological mechanisms of their formation and yet different biological functions. Efficient and accurate identification of different types of repeats is therefore an important bioinformatics problem.

Since 1999, we have been working on different (combinatorial, algorithmic and applicative) issues of tandem repeats (periodicities) in DNA sequences[5]. Developed algorithmic techniques have been implemented in the `mreps` software [56] (see Section 4.1).

As far as distant (interspersed) repeats are concerned, computing them can be regarded as a particular application of the general-purpose local alignment computation. However, this specific application can be seen as a problem on its own, and several programs exist for computing two-copy repeats in genomic sequences (REPUTER, ASSIRC, FORREPEATS and some others). None of those methods is suitable for systematically computing *multi-copy repeats*, i.e. sequences that have multiple (more than two) occurrences in a given genome. Somewhat unexpectedly, this turns out to be a difficult problem (see e.g. [72]) that is important in numerous applications including some projects conducted in our group.

3.2.3. Seed-based protein search

Spaced seeds (see Section 3.2.1) have been applied very successfully to increase the efficiency of DNA similarity search. However, little is known about how suitable spaced seeds are for searching protein sequences ([43] is one of the few papers devoted to this issue). One reason for that is that the identity of amino acids in protein comparison plays a lesser role than the identity of nucleotides in DNA or RNA comparison. On the other hand, the increase of the alphabet size from 4 to 20 implies the decrease of reasonable seed length

(typically, from 9-15 in the nucleotide case to 2-4 in the protein case). This might suggest that the concept of spaced seeds becomes vacuous for the protein case. We believe, however, that this is not the case.

In [60], we proposed a formalism of *subset seeds* that allows one to take into account in a very flexible way complex similarity relations between letters of the sequence alphabet. For example, traditional spaced seeds for the DNA case can only distinguish between nucleotide matches and mismatches, while subset seeds are able to make finer distinctions between different types of mismatches, which brings an additional increase in sensitivity. This approach seems to be particularly suitable for protein sequences, where we have to assign different weights to different pairs of amino acids. Applying the subset seeds approach to the protein case seems very promising but raises new questions. Furthermore, it is very likely that efficient seeding methods for proteins will involve *multiple seeds* rather than single seeds. Designing such seeds is a challenging issue. To sum up, the general problem here is to develop an efficient seeding method for similarity search in protein sequences, including methods for sensitivity and selectivity estimation, seed design and other related problems. Among numerous applications that such a method could have, we mention the mass spectrometry and more precisely the MS/MS technology for protein identification that uses a database search at one of its stages. Improving the performance of this search would bring an important improvement to the whole technology.

3.3. Non-coding RNA analysis

Keywords: *non-coding RNA*, *secondary structure*, *structure alignment*, *structure inference*.

As mentioned in the introduction to this report, we intend to develop sequence analysis tools that are more particularly devoted to the annotation of non-coding regions of the genomes. In this perspective, non-coding RNAs, also known as *RNA genes*, play a major role. They are nucleic acid molecules that are not translated into proteins. Their functions are strongly related to their structure. RNA molecules have the capacity to form isosteric base pairings: Watson-Crick (`A-U` and `G-C`), wobble (`G-U`) or even non canonical pairings. These pairings result in a hierarchical folding that determines the spatial organization of the RNA molecule and its function in the cell (RNA/protein interactions, RNA/RNA interactions etc.). From a combinatorial point of view, RNA is a complex object. It is usually modeled by trees or by graphs.

The study of RNA genes has recently undergone a deep change of perspective caused by the discovery of the essential role of RNA genes in the cell, in the expression regulation, together with the sequencing of full genomes and the availability of an increasing number of families of homologous RNA genes. Non-coding RNAs are now recognized to be essential actors of the eukariotic complexity. There is currently a need for computational tools for a systematic analysis of those genes, analogous to those available for protein-coding genes.

3.3.1. RNA gene prediction

The problem of gene prediction consists in locating non-coding genes in newly sequenced genomes. *Ab initio* prediction is currently an open question. In contrast to protein coding genes, RNA genes lack simple biological signals such as START and STOP codons, or a codon usage bias. Basic questions such as the existence of a nucleotide composition bias or the significance of free energy level are still controversial. Discovering any statistical or information-theoretic characteristics proper to RNA sequences with respect to the background genomic sequence would shed a new light on the properties of RNA genes. Besides intrinsic sequence features, a general paradigm in RNA analysis is that a better prediction accuracy can be reached by employing *comparative analysis* methods (see Section 3.1). The idea is that the structure is preserved by evolution, and mutations observed between homologous RNA sequences should not occur randomly: they are consistent with the formation of base pairs and occur at correlated compensatory positions. The underlying assumption is that RNA genes are characterized by the preservation of their structure through evolution. A conserved structure over divergent sequences suggests that this structure should be functionally important. Under this perspective, gene prediction is partially reduced to the problem of determining if sequences actually share a common structure. We developed recently a CARNAC software for structure prediction [70], [75], [78] (see Section 4.1). But gene prediction raises several new questions. The first one is concerned with the statistical significance of a predicted structure. There are many results about word statistics in genomic sequences, but

these theories have no counterpart for structured motifs such as RNA motifs. The other problem is algorithmic efficiency to allow for a genome-scale annotation.

3.3.2. Structure alignment and motif location

A problem complementary to RNA structure prediction is RNA comparison and RNA pattern matching. It occurs when we know at least one representative structure for the family of homologous RNA genes under consideration. For example, this structure could have been obtained from crystallography experiments or inferred from a phylogenetic analysis. Similar to the usual sequence alignment and sequence pattern matching (see Section 3.2), the goal here is to bring out elements of the structure that have been conserved through evolution and therefore are more likely to be functional. Thus, structural alignment of RNA sequences is a basic operation in RNA analysis, just as the usual sequence alignment is a basic operation in DNA analysis. Comparison of RNA structures should take into account several levels of information corresponding to hierarchical RNA folding: sequence, secondary structure, tertiary interactions. A corresponding model can be represented by labeled ordered trees or arc-annotated sequences. We have a strong experience in working with this type of models [3], [76], [77]. Such models can also be applied to the approximate RNA pattern matching problem, that can be seen as an extension of the alignment problem. Given a description for an RNA family, the goal here is to locate all its potential occurrences on a genomic sequence. Existing methods should compromise between efficiency and sensitivity, and even the fastest programs are not suitable for a genome-scale analysis [50]. These methods rely mainly on probabilistic models of context-free stochastic grammars. There is a lack of pure algorithmic approaches, based on the same combinatorial models as for the structure alignment. Such algorithms could be combined with a probabilistic analysis that would provide a rigorous foundation for the scoring systems. Another line of research for that problem is the indexing of big quantities of RNA data (e.g. RNA databases) in order to perform a fast search of RNA structures. Instead of being based on index data structures designed for sequences, one could index structure elements such as potential stems for example. Designing an efficient index for RNA search would be a major advance for the RNA pattern matching problem.

3.4. Cis-regulatory sequence analysis

Keywords: *cis-regulatory regions*, *phylogenetic footprinting*, *position weight matrices*, *transcription factor binding sites*, *transcription factors*.

Another important aspect of the analysis of non-coding regions in DNA concerns gene regulation. Gene expression in eukaryotic cells is controlled at several levels: mRNA transcription, mRNA processing, protein synthesis, post-translational modifications, RNA degradation. Genome analysis can help to elucidate the very first step in this chain: transcriptional regulation. Transcription of a gene is controlled by regulatory proteins – such as transcription factors (TFs) – that bind to the DNA, mostly in non-coding regions preceding the genes. This protein/DNA interaction requires a binding site whose sequence pattern is more or less specific to each TF. Identification of transcription factor binding sites (TFBSs) is a notoriously difficult task because motifs corresponding to TFBSs have a very low information content: they are usually short (around 5-15 bases) and degenerate. Modeling, identification and analysis of TFBSs is one of major bioinformatics challenges.

3.4.1. Over-represented motif identification

Most successful approaches nowadays integrate two complementary sources of information: statistical over-representation of motifs and conservation of the TFBS across species with phylogenetic footprinting. A way to enhance the specificity of TFBS prediction is to work with a collection of functionally related genes that are believed to be co-regulated, such as groups of genes derived from microarray experiments. In this setting, pattern recognition algorithms can be used to identify overrepresented motifs in the upstream regulatory regions of genes. Numerous tools became available for this problem for the past few years. While there have been several successful applications to different bacteria and low eukaryotes (such as yeast), this task gets much more difficult for higher eukaryotes [74].

The most popular model of TFBSs is given by *Position Weight Matrices* (PWMs), which are probabilistic models of DNA approximate motifs. Databases such as TRANSFAC or JASPAR contain hundreds of curated PWMs for vertebrate organisms. Several recent algorithms address the problem of finding over-represented TFBSs modeled by PWMs [46], [54]. However, the problem is very far from being solved in a satisfactory way and further biologically relevant criteria should be used to enhance the prediction quality. Furthermore, the completion of whole genome sequencing projects for several mammals in near future will provide us with a sufficient number of organisms at the right evolutionary distance in order to perform a phylogenetic footprinting for human data [45]. This research direction is therefore very promising and has still a lot of progress to be made.

3.4.2. Genome scale analysis

As implied by the previous paragraph, the analysis of cis-regulatory regions requires a massive search of motifs in long genomic sequences coming from different species (so called *network level*). This task constitutes then an important computational problem in itself. This *PWM matching problem* includes several lines of research. The basic problem consists in locating all TFBSs for a single PWM. For this purpose, it could be possible to take advantage of topological regularities of PWMs, and of properties of the associated threshold score, following the example of exact pattern matching algorithms. Another algorithmic problem is to locate all occurrences for a large collection of PWMs, such as TRANSFAC combined with JASPAR for example. In this context, the computation can be speeded up considerably by preprocessing the set of PWMs and taking advantage of the mutual content information of the PWMs. Lastly, efficient algorithms for the PWM matching problem could open a way to a systematic exploration of regulatory regions, highlighting cooperation between TFs. Designing appropriate indexes could help to enhance the query performance [80] and would lead to an advanced TFBS retrieval system.

3.5. Nonribosomal peptide synthesis

Keywords: *amino acids*, *nonribosomal peptide synthesis*, *synthetase*.

The central dogma of molecular biology presents the protein synthesis as a transfer of information from DNA to proteins via transcription and translation. Nonribosomal peptide synthesis (NRPS), as its name suggests, is an alternative pathway that allows production of polypeptides other than through the traditional translation mechanism. The peptides are created here by enzymatic complexes called *synthetases* and the resulting peptides are generally short, 2 to 50 residues. NRPS produces several pharmacologically important compounds, including antibiotics and immunosuppressors. This biosynthesis pathway is found in many bacteria and fungi. Recent surveys on that issue appeared in [59], [65].

From a combinatorial viewpoint, peptides produced by NRPS show peculiar features compared to traditional proteins. First, they can contain standard as well as non-standard amino acids. Secondly, amino acids are linked not only by an amino-peptide link, but also by non-conventional links that form a non-linear peptide backbone. There exist iterative and nonlinear NRPS configurations that generate more complicated structures. Consequently, some peptides form cycles, unusual branching or repeats leading to various topological structures. Very few computational tools exist today for dealing with such peptides (encoding, comparing, searching, ...). NRPS-PKS [41] is one of them that is mostly devoted to the analysis of synthetases and enzymes associated to the production process and does not include features to handle nonribosomal peptides.

Our project here is to develop a comprehensive computational tool, called NORINE, to work with nonribosomal peptides. One goal of NORINE is to be a complete database of annotated NRPS peptides. Another goal is to allow a biologist to compare NRPS molecules according to different criteria, as well as to search through them for a given pattern. The latter brings up non-trivial computational problems of graph processing.

This work is done in collaboration with Lille-based biologists (see Section 7.1).

3.6. General models and tools

Keywords: *discrete algorithms*, *discrete probability*, *high-performance computing*, *statistics*.

In contrast to Sections 3.2-3.5, this Section does not present a specific research area but rather three major groups of tools that we use in our research. We highlight here three themes that are applied to virtually all above-mentioned research projects. These are *discrete algorithms* on the one hand, that constitute a major foundation of the project, and *statistics* and *high-performance computing* on the other hand, that are rich external resources for us. Note that these three tools are of different nature but, on the other hand, are common to most of the problems described in Sections 3.2-3.5.

3.6.1. Discrete algorithms

3.6.1.1. Combinatorial algorithms

The scientific core of our work is the design of efficient algorithms for the analysis of biological macromolecules modeled by combinatorial objects. Indeed, biological macromolecules are naturally and faithfully modeled by various types of discrete structures: string for DNA, RNA and proteins, trees and graphs for RNA and proteins. Furthermore, computational biology applications lead to the emergence of new combinatorial instances for these structures: spaced seeds for sequence analysis, arc-annotated sequences or 2-interval graphs for RNA structures, profiles for PWMs, ... Thus, this "interaction" is a mutual enrichment.

Building rigorous mathematical models is an important primary goal of our project. To such models, we apply the whole large spectrum of algorithmic techniques that has been developed in the area of discrete algorithms during last decades and develop new algorithmic methods when necessary. The area of string algorithms (sometimes termed *stringology*) continues to be a very active area of research. Graph and tree algorithms have been at the heart of computer science for decades.

Using combinatorial data structures has an advantage to provide a formal way to measure the efficiency via the notion of algorithmic complexity. We systematically apply the complexity analysis to our algorithms in order to improve their performance, both in terms of time and space requirements. Efficiency may be a critical point for algorithms dealing with large data sets. Moreover, many real-life bioinformatics problems are intrinsically difficult (often NP-complete or harder): multiple alignment, sensitivity of a set of seeds, comparison of RNA structures with expressive models, etc. We need to develop heuristics that nevertheless *guarantee* certain performance characteristics, relevant to the underlying biological problem.

3.6.1.2. Indexing techniques

Discrete structures are intimately related to powerful *indexing* structures that allow a data set to be stored and queried efficiently. Indexing structures are widely-used in computational biology as they are particularly interesting for the analysis of genomic data. As an example, virtually all similarity search program (see Section 3.2) use an index for storing seed keys. Indexing problems appear in RNA matching (as mentioned in Section 3.3) as well as in PWM search (Section 3.4). Thus, designing efficient index structures is crucial for many of our research topics and holds therefore a particular place within the scope of our studies.

3.6.2. Statistics and discrete probability

When dealing with large input data sets, it is essential to be able to discriminate between noisy features observed by chance from those that are biologically relevant. The aim here is to introduce a probabilistic model and to use sound statistical methods to assess the significance of some observations about these data, e.g. of the output of a software program. Examples of such observations are the length of a repeated region, the number of occurrences of an approximate motif (DNA or RNA), the free energy of a conserved RNA secondary structure, the score quality of a motif specified by a PWM, the overlapping rate of two motifs, ... The fundamental underlying idea here is that only statistically significant (low-probability) observations (with respect to an appropriate probabilistic model) can potentially correspond to a biological meaning.

Another important situation in our work where the probabilistic analysis comes into play is related to the algorithmic complexity issue. As we noted above, when the algorithmic complexity of a problem is too high, we need to develop non-exhaustive methods that guarantee some performance characteristics. One way of doing this is to ensure that while our method does not verify the requirements on *all* data, the fraction of missed results is *statistically small* with respect to a given probabilistic model.

3.6.3. High-performance computing

Using high-performance computing techniques and facilities is a necessity for our project, due to high volumes of genomic data that we often have to deal with. Therefore, high-performance computing is an additional technological tool that we use to achieve our goals.

We are in contact with the DOLPHIN project-team that is the promoter of the GRID 5000 farm in Lille. We were regular users of the GRID 5000 farm and part of the local GRID 5000 community. So far, it allowed us to reduce considerably the CPU time for our tests and large scale validations. For example, it allowed us to carry out an exhaustive analysis of large public databases of coding, non-coding and unannotated conserved sequences (Pandit, RFAM, UCSC genome browser) with the caRNAc program enriched by a coding model (see Section 3.3).

Another way to enhance computing performances is to use *specialized computer architectures* to obtain a fine-grained parallelism [7]. We collaborate with the SYMBIOSE project-team (INRIA-Rennes) that builds prototypes designed to index large amounts of data (see Section 7.2). More generally, we are intersted in the new *massively multicore architectures*. The graphic processing units (GPU) are a first step toward those architectures, and we began in 2008 to conceive algorithms for some parallel applications on the GPU. We plan to further pursue this line of research in the following years.

4. Software

4.1. Introduction

Software development is an important part of our work as many of the algorithmic techniques we develop are implemented in deliverable software. We maintain a server accessible via http://bioinfo.lifl.fr/ for distributing our software and executing it through web interfaces.

In 2008, we delivered a new software, called MAGNOLIA, for advanced multiple sequence alignment. We also present other software programs developed in the team and that are still under active development.

4.2. MAGNOLIA

Keywords: *multiple sequence alignment*, *non-coding RNA*, *protein coding sequences*, *structure prediction*.

Participants: Arnaud Fontaine, Antoine de Monte, Hélène Touzet.

URL: http://bioinfo.lifl.fr/magnolia/

MAGNOLIA is a new software for multiple sequence alignment that exploits our ideas coming from comparative analysis presented in Section 5.4.3. It takes as input a set of unaligned nucleic acids sequences, classifies the sequences either as coding RNAs or non-coding RNAs and produces a multiple sequence alignment based on the the appropriate evolutionary pattern. When sequences are predicted as coding, then the multiple alignment relies ont he putative amino-acid sequences. When sequences are predicted as non-coding, then the multiple alignment relies on the putative conserved secondary structure.

4.3. YASS suite

Keywords: *homology*, *sequence alignment*, *sequence similarity*, *subset seeds*, *transition constrained seeds*.

Participants: Laurent Noé, Antoine de Monte.

URL: http://bioinfo.lifl.fr/yass

YASS [69] [9] is a software for computing similarity regions in genomic sequences (local alignment). The first version of YASS has been released in January 2003. From the algorithmic point of view, YASS is based on two main innovations that insure a high sensitivity of the search: one is a powerful seed model, called *transition-constrained seeds*, that extends the basic spaced seed paradigm (Section 3.2), and the other is a new *hit criterion* that specifies the way that the seeds are used to detect potential similarity regions. IEDERA is an accompanying software that implements the work of [61]. This year, we delivered release v1.14 of YASS that mainly improves 64-bit and multithreading support.

4.4. Noncoding RNA tools

Keywords: *non-coding RNA*, *structure comparison*, *structure inference*, *structure prediction*.

Participants: Arnaud Fontaine, Antoine de Monte, Hélène Touzet.

URL: http://bioinfo.lifl.fr/RNA

On the subject of RNA analysis, CARNAC is a program for RNA structure prediction. The software is based on a multicriteria approach combining thermodynamic stability and phylogenetic information. Its implementation is based on dynamic programming and graph theory methods. CARNAC has proved to be particularly efficient on large and noisy data sets [52], and is presented in a book chapter devoted to comparative genomics [78]. This year, CARNAC has undergone a major update, described in Section 5.2.2. GARDENIA is a complementary tool for comparing and aligning RNA structures, taking into account both the sequence and the structural information. It is based on the paradigm of the optimal common superstructure, that was introduced in [13]. GARDENIA appears to be more robust than similar existing programs, such as those of the Vienna Package.

4.5. TFM suite

Keywords: *cis-regulatory regions*, *phylogenetic footprinting*, *position weight matrices*, *transcription factor binding sites*, *transcription factors*.

Participants: Aude Liefooghe, Hélène Touzet, Jean-Stéphane Varré.

URL: http://bioinfo.lifl.fr/TFM

Our research on cis-regulatory regions described in Section 3.4 is being implemented in a series of programs devoted to the location and processing of Position Weight Matrices. This platform includes currently three programs. The TFM-EXPLORER software is dedicated to the inference of locally over-represented motifs in mammalian genomes [1]. TFM-Explorer has been released in August 2006, and has been used by several biology research groups [81], [67], [79]. The TFM-Scan program implements efficient algorithms for the location of PWM matrices on a sequence [8], [24]. TFM-Pvalue is a program to compute score thresholds for PWMs [10].

4.6. Protea

Keywords: *coding sequence identification*, *exon prediction*.

Participants: Arnaud Fontaine, Hélène Touzet.

URL: http://bioinfo.lifl.fr/protea

PROTEA is a new software for identifying evolutionary conserved coding sequences using a comparative analysis of genomic sequences. It relies on ideas presented in Section 5.4.1. PROTEA takes as input a set of unaligned similar sequences and classifies this set into coding or other sequences. As a byproduct, it builds a multiple sequence alignment based on the putative amino acid sequences according to the predicted reading frame.

4.7. Norine

Keywords: *database*, *nonribosomal peptide synthesis*.

Participants: Ségolène Caboche, Gregory Kucherov, Maude Pupin.

URL: http://bioinfo.lifl.fr/norine

We continue to develop a database of NRPS peptides called NORINE[1]. This is a unique resource as there has been no centralized depository of these data before. Among existing related resources, NRPS-PKS[2] is focused on the synthases and contains only a very limited number of peptides, other resources like PubChem[3] or ChEBI[4] have a much more general scope and are not devoted to NRPS peptides. Note that each entry of NORINE is generally obtained from the literature and is manually curated. The database is freely accessible through the Web. The entries contain various annotations of the peptides: names and synonyms, biological activities, "monomeric" structure, chemical composition, molecular weight, producing organism, bibliography references, possible links to others databases such as PubChem or UniProt. The user can query the annotations and the structures via a web interface in order to select the NRPS peptides that correspond to different search criteria.

This year, NORINE contains more than 1000 peptides and updated annotations. The data dedicated to the amino acids was curated and annotations were added such as SMILES representation.

5. New Results

5.1. Sequence similarity and repetitions

Keywords: *high-performance computing*, *homology*, *repeat*, *sequence alignment*, *sequence similarity*.

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

5.1.1. Estimation of seed sensitivity

Following our previous work in which we proposed and studied the idea of *subset seeds* for sequence comparison [60], this year we studied the *subset seed automaton* which plays a central role in the algorithm to estimate the performance (sensitivity) of those seeds. This work has been presented to the 12th International Conference on Implementation and Application of Automata (CIAA 2007) [61]. The main novel contribution of this work is an efficient incremental linear-time algorithm to construct the subset seed automata. It is important to note that this automaton can be generalized to other pattern matching problems, such as matching of sequences over an alphabet including ambiguous letters. Note that the automaton is implemented in the IEDERA software (see Section 4.3). An extended journal version of this paper is still under submission to a journal.

5.1.2. Seeds for protein search

This year we continued our work on seed-based comparison of protein sequences. Its main motivation has been to apply to protein sequences the concept of *subset seeds* proposed in [60] for DNA sequences. We studied several approaches to the design of a *seed alphabet*, which is an important preliminary step to constructing efficient seeds. Both *non-transitive* and *transitive* alphabets have been studied. For transitive alphabets, we studied two different approaches, based on either a pre-defined hierarchical tree of amino acids (such as those proposed in [66], [62]), or on specially designed amino acid hierarchies that take into account foreground and background distributions of amino acids in target protein sequences.

Seeds over designed alphabets have been tested on probabilistic models as well as on real data. It turns out that their performance (selectivity/sensitivity ratio) is comparable to (or even, in certain cases, better than) that of BLAST. This result is interesting as the formalism of subset seeds is weaker than the one of BLAST, which allows a more simple and more efficient implementation. The latter feature has been used in our work on efficient hardware implementation of those seeds, described in the next section.

[1] **no**n-**ri**bosomal peptides, with **ine** as a typical ending of names of nonribosomal peptides
[2] http://www.nii.res.in/nrps-pks.html
[3] http://pubchem.ncbi.nlm.nih.gov
[4] http://www.ebi.ac.uk/chebi

A first conference paper describing these studies has been published this summer [26]. The extended and complete journal version is under submission.

5.1.3. Neighborhood indexing

Within the 2006-07 ARC Flash collaboration with the SYMBIOSE team in INRIA-Rennes (see section 7.2), we designed in 2007 a technology that implements subset-seed-based search for protein sequences (see previous section) in a specialized parallel hardware [71].

In 2008, we studied the consequence of reducing the amino acid alphabet in the case of protein similarity searches. We showed that an optimal neighborhood indexing combining an alphabet reduction and a longer neighborhood leads to a reduction of 35% of memory involved into the process, without sacrificing the quality of results nor the computational time. This approach led us to develop a new kind of substitution score matrices and their associated e-value parameters. In contrast to usual matrices, these matrices are rectangular since they compare amino acid groups from different alphabets. The website http://bioinfo.lifl.fr/reblosum proposes a selection of such matrices as well as an interface to compute other matrices. A journal article with those results has been accepted [17].

5.1.4. Runs and palindromes

We continued algorithmic studies on palindromic and periodic structures in words (sequences). In [23], we studied the problem of efficient computation of *gapped palindromes*. More specifically, we defined two natural classes of gapped palindromes, named *long-armed* and *length-constrained* palindromes. For each of these two classes, we proposed an efficient algorithm for computing all palindromes in time $O(n + S)$, where n is the sequence length and S the number of output palindromes. The algorithms are based on advanced string processing techniques (longest extension functions, reversed Lempel-Ziv factorization, dynamic data structures). It is important to note that both algorithms extend to biological palindromes, that makes them very useful in identification of RNA structures in genomic sequences.

Moreover, we studied the problem of the maximal number of runs in a string. It was known since 1998 than this number $\rho(n)$ is linear in the length n of the string [57]. Lower bounds and upper bounds have been recently provided by different teams. However, very few properties were known for the $\rho(n)/n$ function. In [20], we bring some improvements on the problem of the limit of the maximal numbers of run, showing that this limit exists and is never reached.

5.2. RNA genes and RNA structures

Keywords: *RNA*, *base pairings*, *secondary structure*, *structure alignment*, *structure inference*.
Participants: Arnaud Fontaine, Mathieu Giraud, Antoine de Monte, Azadeh Saffarian, Hélène Touzet.

5.2.1. RNA structure comparison

We continued our work on the RNA alignment hierarchy, originally initiated in [42]. This alignment hierarchy provides a general unifying framework to express the comparison of RNA structures represented by specific graphs, called arc-annotated sequences. It encompasses main existing models, such as tree edit distance, general edit distance, tree alignment. We carried out experimental analyses of the average complexity of some polynomial instances of the hierarchy [13].

In the context of Brasero ANR, we also took part to a multidisciplinary working group devoted to benchmarking RNA secondary structure comparison algorithms. A preliminary release of this work was presented in [27].

5.2.2. RNA structure prediction

CARNAC is a software for RNA structure prediction that has been developed in the team since 2003. This year, we added several essential improvements to the method. First, we modified the core algorithm of the folding step, resulting in a large speed up. Then we extended the evolutionary model to take into account similar sequences with low evolutionnary distance. For that, we proposed a novel approach that combines the multiple sequence alignment oriented paradigm with the Sankoff paradigm. The goal is to achieve higher sensitivity in the prediction. A publication on this subject is in preparation.

5.2.3. RNA suboptimal structures

We started at the end of 2007 a new project to have better data models for the set of all secondary structures, including the suboptimal ones, of a given RNA. We studied some properties of a graph encoding those structures, and proposed algorithms for the search of saturated secondary structures. An article on this subject is in preparation.

5.2.4. Evolution of ribosomal RNAs in echinoderms

This research is a collaborative work with DIMAR Lab (Marseille). DIMAR Lab collected a large set of sequences for the D8 domain of ribosomal RNAs in echinoderms, which are marin invertebrate animals such as sea urchins. D8 domain is of high importance to understand the evolution and the phylogeny of these species. In this context, our contribution consisted in analysing the secondary structure of the domain [14].

5.3. Cis-regulatory sequence analysis

Keywords: *cis-regulatory regions*, *phylogenetic footprinting*, *position weight matrices*, *transcription factor binding sites*, *transcription factors*.

Participants: Jean-Stéphane Varré, Hélène Touzet, Aude Liefooghe, Mathieu Giraud.

5.3.1. Single PWM matching problem.

In 2006 we produced a method able to efficiently looking for occurrences of a set of PWMs at a genome scale. Methods addressing this problem were missing. This year, we addressed the problem of efficiently finding occurrences of a single PWM. This problem has recently attracted some interest. It can be viewed as a special case of the exact multiple pattern matching problem with a very high number of patterns. Methods based on the building of a data structure such as Finite State Automata cannot be used because of the huge memory space needed to store it. We proposed to take advantage of the non-overlapping property of PWMs to extend algorithms that use this property in exact pattern mathing. We then proposed an extension of the Knuth-Morris-Pratt algorithm that allowed to achieve a speedup of three for the searching phase while keeping a reasonnable preprocessing time [24].

5.3.2. Module identification and matching.

We began to investigate the problem of discovering cis-regulatory modules. A module is a set of transcription factors interacting for the regulation of a gene. Modules can be detected because we can observe co-occurrences of transcription factor binding sites in the promotor region of a gene given a set of genes with the same function from several species. Such observations show that the distance between the binding sites are often constrained. Two directions of research have been initiated. The first one relates to the expansion of TFM-Explorer (see 4.5) in order to detect modules. The second one is related to module matching. Given a module, defined by a set of PWMs and distances between them, how to efficiently search for occurrences over a genome ?

5.4. Comparative genomics applications

Participants: Arnaud Fontaine, Mathieu Giraud, Benjamin Grenier-Boley, Antoine de Monte, Laurent Noé, Hélène Touzet.

5.4.1. 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 [16].

5.4.2. RNA gene prediction through seed-based comparative genomics

As mentioned previously, sequence comparison is widely used to help to discover new non-coding RNAs in newly sequenced genomes. In this perspective, we started to compare and to evaluate different similarity search heuristics: usual BLAST contiguous seeds and YASS multiple spaced seeds. RNA gene identification is a difficult task as the level of conservation between RNA genes tends to be lower than for coding genes. Spaced seed-based approaches show a higher sensitivity than contiguous seeds. Furthermore, we designed optimized spaced seeds on the non-coding RNA RFAM database [53] and estimated their theoretical sensitivities. We discovered some bias in the benchmarks of [51]. Finally, following the methodology of [73], we compared the predictions on non-coding RNA candidates versus known RNA on *E.coli*. This work was presented in [49]. We are preparing a paper on those themes.

5.4.3. Multiple alignment via comparative analysis

We have proposed a new method to construct multiple alignments of nucleic acid sequences. These sequences are recognized to be hard to align because similarity is often reduced at the DNA level. Regarding protein coding genes, nucleic acid sequences exhibit a much larger sequence heterogeneity compared to their encoded amino acid sequences due to the redundancy of the genetic code. The same situation holds for non-coding RNA genes. The spatial structure evolves slower than its primary structure. Our idea is to take into account the putative function of the sequences and to incorporate this functional information into the alignment. The algorithm is based upon the comparative paradigm: it extracts information from the similarities and differences in the data, and searches for a specific evolutionary pattern between sequences before aligning them. This has been implemented in a software named MAGNOLIA and evaluated on large experimental data sets [15].

5.5. Nonribosomal peptide synthesis

Keywords: *amino acids*, *nonribosomal peptide synthesis*, *synthetase*.

Participants: Ségolène Caboche, Gregory Kucherov, Maude Pupin.

As presented in Section 4.7, NORINE is the first centralized resource exclusively devoted to storing and manipulating (retrieving, comparing, searching, ...) nonribosomal peptides. Note that the number of known such peptides is counted by hundreds and is still growing. Note also that these peptides have a very diverse structure: they can be linear, branched, totally cycled, cycled with branches and double or tri-cycled. In contrast to “conventional” proteins that are composed of 20 different amino acids, nonribosomal peptides can contain more than 400 different monomers (amino acids and other molecules). Finally, they have several important activities, such as antibiotic, anti-inflammatory, antithrombotic, antitumor, calmodulin antagonist, immunomodulating, protease inhibitor, siderophore, surfactant, and toxin.

Since last year, the search for all molecules containing a given *structural pattern* is available through NORINE website. An article to present the algorithm behind this search was written and submitted to BMC Structural Biology. It is under review. A huge work was done to update the data stored in NORINE : new peptides were added, the annotation of the "old" ones was updated, the data on the monomers was extended. We also provide new tools on the website to give the opportunity to the users to submit annotations on a peptide already stored in NORINE or to submit a new peptide.

5.6. Other new results

Participants: Aude Darracq, Jesper Jansson, Sylvain Guillemot, Alban Mancheron, Jean-Stéphane Varré.

We worked on algorithms and combinatorics related to the construction of phylogenetic trees and phylogenetic networks [22], [21], [19]. An other subject was the inference of frequent itemsets [30], [29]. Finally, we obtained first results in the study of genomic rearrangements in the beet mitochondrial genome (poster [39], an article is in submission).

6. Contracts and Grants with Industry

6.1. NVIDIA

We began this year some contacts with NVIDIA, one of the leading companies in producing graphics processing units (GPUs). The CUDA librairies, released in 2007, abstract and simplify the development on those GPUs. We asked support from NVIDIA University Relations, and NVIDIA gave to the team a Tesla S870 computing server (rack 1U with 4 GPUs) to test our parallel algorithms.

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.

- We are a part of the *Génopole de Lille*, with our software available through the *Génopole* website[5].
- The project on *nonribosomal peptide synthesis* is based on a collaboration with the laboratory ProBioGEM (*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 is co-supervised by Valérie Leclère from ProBioGem. A new PhD work is starting on this subject: 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 beet mitochondrial genome. 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 will be acquired thanks to a Genoscope project (accepted). A PhD student (Aude Darracq) is co-supervised on this subject.
- We are associate members of the research federation *IRI* (Interdisciplinary Research Institute – USR CNRS, This institute is designed to foster interactions between biologists, computer scientists, mathematicians, physicists, chemists and engineers on topics related to the structure, dynamics and robustness of regulatory networks.
- 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.
- We continue the collaboration with F. Sebbane (INSERM U 801) on the analysis of *Yersinia pestis* genome for the discovery of small non-coding RNAs.

[5] http://www.genopole-lille.fr

7.2. National initiatives and cooperations

7.2.1. National initiatives

We participate in the following national projects:

- ANR BRASERO (Biologically Relevant Algorithms and Software for Efficient RNA Structure Comparison), *Programme blanc 2006*. The project aims at providing relevant and efficient tools for the RNA comparison problem. Other participants : LRI (University Paris Sud), LaBRI (University Bordeaux 1), Helix (INRIA Rhône-Alpes).
- ANR COCOGEN (Comparaison of Complete Genomes), *Programme blanc 2007* L.Noé together with MAB team of LIRMM (Montpellier), MIG and UBLO team of INRA (Jouy en Josas), INA-PG (Paris).
- inter-Genopole project *NCRNA: Non-Coding RNAs*, funded by RNG-Renabi (2007-09). This project involves the bioinformatics platforms of Génopole Toulouse-Midi-Pyrénées and Génopole Nord Pas-de-Calais, and is supervised by C. Gaspin (Toulouse-Midi-Pyrénées). The objective is to develop in a concerted way an open-source integrated platform allowing in silico ncRNA gene annotation in genomic sequences.
- working groups *Sequence analysis* and *Structural bioinformatics* of the multidisciplinary *GDR Molecular bioinformatics*[6].
- working group *Combinatoire des mots, algorithmique du texte et du génome* of the *GDR Informatique Mathématique*[7].

7.2.2. National cooperations

- University Marne-la-Vallée – Institut Gaspard Monge, with G. Blin, RNA comparison, (H. Touzet)
- University Paris-Sud – LRI, with A. Denise, RNA comparison, (H. Touzet)
- Evry, Laboratoire Statistique et Génome, with E. Corel, C. Devauchelle, A. Grossman, A. Hénaut and I. Laprevotte, alignment-free sequence comparison (M. Pupin)
- Institut de Mathématiques de Luminy, with G. Didier, alignment-free sequence comparison (M. Pupin)
- The following french scientists were invited in the past year to give a talk at the team seminar: D. Gautheret (Univ. Paris Sud), M. Zytnicki (INRA Toulouse), A. Ouangraoua (Univ. Bordeaux 1), A. Labarre (Univ. Bruxelles), D. Hot (Institut Pasteur de Lille)
- UR895 Génétique Microbienne (INRA Jouy-en-Josas), with J.-M. Batto and S. D. Ehrlich, GPU parallelisation of algorithms with new sequencers (M. Giraud), new collaboration started this year
- Rennes, Symbiose team (Univ. Rennes 1 / INRIA Rennes), with D. Lavenier and P. Peterlongo. After the INRIA 2006-07 *Action de Recherche Coopérative (ARC) "Optimisation de graines et indexation des banques d'ADN sur mémoire FLASH reconfigurable"*, we proposed in 2008 a new comparison method for protein on a reduced alphabets (see Section 5.1.3).
- DIMAR – Diversité, évolution et écologie fonctionnelle marine UMR 6540, Université de Marseille, with A. Chenuil (H. Touzet)

7.3. International initiatives and cooperations

7.3.1. European projects

The proposal, called NOVAPIC (*Novel Assembly Line Catalytic Machinery for Effective Production of Innovative Bio-active Compounds*), of a large collaborative European project within the call *Food, Agriculture, Fisheries and Biotechnologies* of FP7 (call KBBE-2007-2A) submitted last year has been selected at the first evaluation stage and obtain a score of 12/15 at the second stage. Unfortunately, this was not enough to be funded. Our role in this project was to provide some bio-informatics tools to study non-ribosomal peptides and their synthetases.

[6] http://www.gdr-bim.u-psud.fr
[7] http://www.gdr-im.fr/

The subject on non-ribosomal peptides is part of two submitted proposals. A proposal called PHYTOBIO (*Développement et promotion de nouveaux produits phytosanitaires pour la lutte biologique contre les maladies des plantes*) submitted to *INTERREG IV, Coopération territoriale européenne, France - Wallonie - Vlaanderen* and a *Marie Curie Initial Training Networks (ITN)* of FP7 (call FP7-PEOPLE-ITN-2008) called LIPOCONTROL *Engineering of novel lipopeptides for plant pathogen control*.

7.3.2. Foreign visitors

- Professor Tetsuo Shibuya, from the Human Genome Center of the University of Tokyo visited our group for two weeks in February 2008 and gave a talk at the team seminar.
- Peter Steffen, from University Bielefeld (Germany), visited our group for three days in October 2008, and gave a talk in the LIFL seminar. He collaborates with M. Giraud on a GPU implementation for the ADP (algebraic dynamic programming) methodology.

7.3.3. Bilateral cooperations

- Germany, Bielefeld University, R. Giegerich, P. Steffen: GPU parallelisation of ADP (Algebraic Dynamic Programming) methodology (M. Giraud, visit in August 2008), new collaboration started this year
- Poland, Warsaw University, A. Gambin, S. Lasota: seed-based search in protein sequences (G. Kucherov, L. Noé),
- UK, Cambridge, Isaac Newton Institute for Mathematical Sciences, with C. Semple: phylogenetics (S. Guillemot)
- UK, London, King's College, with K. Iliopoulos, M. Crochemore: string processing (G. Kucherov)
- Brooklyn College, CUNY, with Prof. Dina Sokol: joint work (G. Kucherov)
- Russia, Moscow University, with R. Kolpakov: combinatorics of repetitions in words, tandem repeats in DNA sequences and `mreps` software (G. Kucherov)
- Russia, Institute of Mathematical Problems in Biology in Puschino, with M. Roytberg: seed-based similarity search (G. Kucherov, L. Noé)

8. Dissemination

8.1. Organization of workshops and seminars

8.1.1. JOBIM 2008

Sequoia and Dolphin teams organized the French national bioinformatics conference [8] on june 29-july 3 (supervision of the organizing committee: H. Touzet and L. Jourdan). This major multidisciplinary meeting gathered 350 researchers coming from computer science, biology, physics and mathematics. It included 36 oral communications, 100 posters and 19 software demos. There were six international invited speakers: J.J. Cassiman (Katholieke Universiteit Leuven), J. Demongeot (IMAG), R. Durbin (Wellcome trust Sanger Institute, Cambridge), L. Duret (Biométrie et Biologie Evolutive, Lyon), M. Sternberg (Imperial College, Londres), O. Troyanskaya (Princeton University). The conference was also accompanied by five satellite meetings on July 4 (150 participants).

8.1.2. GTGC working group

J.-S. Varré is one of the committee members of the national GTGC working group[9] (Comparative Genomics Working Group) created in 2005. The group organizes one or two seminar sessions per year on comparative genomics. A large number of presentations are devoted to biological problems. In 2008, the seminar held in Lille, as a satellite meeting of the JOBIM conference.

[8] http://www.lifl.fr/jobim2008
[9] http://biomserv.univ-lyon1.fr/~tannier/GTGC/

8.1.3. PPF Bioinformatique meeting

M. Pupin organized the second one-day meeting for the *PPF Bioinformatique of Lille*, on the 11th of June. Around 50 scientists attend this event.

8.1.4. INRIA Lille GPGPU working group

With J. Allard (EPI Alcove), M. Giraud organizes since September 2008 a weekly working group on "general purpose computing on GPUs". This working group gathers 5-10 scientists from 4 differents teams.

8.1.5. Journées au vert

On January 14-15, 2008, we organized a team two-days seminar in Arras (Pas-de-Calais) in order to discuss current and future research projects carried out in the group.

8.1.6. A journey through term rewriting and lambda-calculi

G.Kucherov was one of the organizers of the one-day workshop *A journey through term rewriting and lambda-calculi* held on May 29, 2008 at Loria in honor of Pierre Lescanne.

8.2. Editorial and reviewing activities

- Editorial Board of BMC Algorithms for Molecular Biology (G. Kucherov)
- Program committee of JOBIM 2008 (M. Pupin, J.-S. Varré), CPM 2008 (G. Kucherov), CSR 2008 (G. Kucherov), PSI 2009 (G. Kucherov), CPM 2009 (G. Kucherov, H. Touzet)
- Reviewer for the journals Algorithmica (J. Jansson), Bioinformatics (L. Noé), BMC Bioinformatics (H. Touzet), Journal of Computer and System Sciences (G. Kucherov), Journal of Bioinformatics and Computational Biology (J. Jansson), Journal of Biotechnology (M. Pupin) Journal of Computer and System Science (G. Kucherov), Journal of Theoretical Biology (M. Giraud), Nordic Journal of Computing (J. Jansson), Nucleic Acids Research (H. Touzet, S. Caboche), Theoretical Computer Science (G. Kucherov), Theory of Computing Systems (J. Jansson) IEEE Transactions on Bioinformatics and Computational Biology (H. Touzet)
- Reviewer for the conferences COCOON 2008 (J. Jannson), CPM 2008 (M. Giraud, J. Jansson, G. Kucherov), JOBIM 2008 (M. Giraud, M. Pupin, J.-S. Varré), MFCS 2008 (M. Giraud, J.-S. Varré), ReConFig 2008 (M. Giraud), STACS 2008 (H. Touzet), SWAT 2008 (J. Jannson), SODA 2009 (G. Kucherov), STACS 2009 (M. Giraud, G. Kucherov).
- Reviewer for American Mathematical Society (AMS)'s Mathematical Reviews (MR) (J. Jansson, three reviews)

8.3. Miscellaneous activities

- Jury of the PhD theses of Goulven Kerbellec (H. Touzet, rapporteur), Olivia Jardin-Mathé (M. Pupin, examinateur), Stefan Canzar (G. Kucherov, rapporteur)
- Reviewers for the french ministry program ANR (G. Kucherov)
- Reviewers for the INRIA "Équipes Associées" program (M. Giraud, G. Kucherov, J.-S. Varré)

8.4. Meetings attended and talks

8.4.1. International Conferences

- LATA 2008, *Language and Automata Theory and Applications*, Tarragona, Spain, April 2008 (M. Giraud [20])
- IWPEC 2008, *International Workshop on Exact and Parameterized Computation*, Victoria, Canada, May 2008 (S. Guillemot [22], [21])

- CSR 2008, *Computer Science in Russia*, Moscow, Russia, June 2008 (G. Kucherov)
- CPM 2008, *Combinatorial Pattern Matching*, Pisa, Italy, June 2008 (G. Kucherov)
- ALBIO 2008, *Algorithms in Molecular Biology*, Vienna, Austria, August 2008 (L. Noé [26])
- *Journées Montoises d'Informatique Théorique*, Mons, Belgium, August 2008 (M. Giraud, G. Kucherov)
- EMBnet Conference 2008, *Leading applications and technologies in Bioinformatics*, Martina Franca, Italy, September 2008 (S. Caboche [37], M. Pupin)
- ECCB 2008, *European conference in Computational Biology*, Cagliari, Italy, September 2008 (S. Caboche [38], M. Pupin)
- WABI 2008, *Workshop on Algorithms in Bioinformatics* Karlsruhe, Germany, September 2008 (A. Saffarian, M. Girdea)
- RECOMB CG 2008, *RECOMB Comparative Genomics*, Paris, France, October 2008 (L. Noé, J.-S. Varré)

8.4.2. National Conferences

- EGC 2008, *Extraction et Gestion des Connaissances*, Sophia Antipolis, January 2008 (A. Mancheron [29], [30])
- JOBIM 2008, *Journées Ouvertes Biologie Mathématique Informatique Biologie*, Lille, July 2008 (S. Caboche [36], A. Fontaine, M. Giraud, M. Girdea [28], B. Grenier-Boley, G. Kucherov, A. Liefooghe, L. Noé, M. Pupin, H. Touzet, J.-S. Varré)

8.4.3. Talks, meetings, seminars

- *Asymptotic behaviour of the number of runs*, London Stringology Days (LSD), February 2008 (M. Giraud)
- *Searching for gapped palindromes*, London Stringology Days (LSD), February 2008 (G. Kucherov)
- *Pipeline d'annotation par génomique comparative*, Séminaire Arena-Renabi, Toulouse, February 2008 (B. Grenier-Boley)
- *Prédiction de structure avec caRNAc: existant et développements en cours*, Séminaire Arena-Renabi, Toulouse, February 2008 (A. Fontaine)
- *Recherche de motifs ARN avec filtrage*, Séminaire Arena-Renabi, Toulouse, February 2008 (H. Touzet)
- *Outils bioinformatiques pour étudier les peptides non ribosomiaux*, Séminaire BIL (Bio-Informatique Ligérienne), Nantes, March 2008 (M. Pupin)
- *Challenges in high-performance bioinformatics computations*, Bielefeld University, August 2008 (M. Giraud)
- *Seeds for biological sequence comparison : an example of how finite automata contribute to genomic studies*, invited talk to *Journées Montoises d'Informatique Théorique*, Mons, Belgium, August 2008 (G. Kucherov)
- *Exact limits on the number of some microruns*, Journées JORCAD, Rouen, September 2008 (M. Giraud)
- *Bioinformatique et calcul haute-performance*, Séminaire Aristote, École Polytechnique, October 2008 (M. Giraud)
- A. Mancheron gived talks in Bordeaux, Rennes, Montpellier, Bruxelles and Glasgow.

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), 17h (M. Giraud, L. Noé, M. Pupin, J.-S. Varré)
- *Computational biology*, master in computer science (M1), 50h (H. Touzet, S. Caboche, together with C. Abbadie)
- *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), 40h (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)
- *Probability and Statistics*, second year of bachelor, 18h (A. Liefooghe)
- *Programming (Pascal)*, first year of bachelor, 36h (M. Pupin, L. Noé)
- *Algorithmics*, third year of bachelor, 25h (A. Liefooghe)
- *Programming (Ocaml, Prolog)*, third year of bachelor, 48h (L. Noé)
- *Networks*, third year of bachelor, 36h (L. Noé)
- *Software project*, third year of bachelor, 35h (J.-S. Varré)
- *Business intelligence*, first year of master, 35h (A. Liefooghe)
- *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

- Lund University, Sweden, *DATN11: Computational Biology*, 2 lectures (J. Jansson)
- *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)
- Coordinator of the Working group *Combinatoire des mots, algorithmique du texte et du génome* of the *GDR Informatique Mathématique* (G. Kucherov, till August 2008)
- Member of the LIFL Laboratory council (H. Touzet)
- Head of PPF bioinformatics – University Lille 1 (H. Touzet)
- Members of the *Commission des Spécialistes* of the University Lille 1 (H. Touzet and J-S. Varré)
- Member of hiring committee (*jury d'audition*) 2008 of INRIA-Rennes - Bretagne Atlantique (G. Kucherov)
- Member of the GTAI INRIA committee (H. Touzet)
- Member of the INRIA evaluation commitee (M. Giraud)

- Member of the INRIA-LNE center commitee (J.-S. Varré)

9. Bibliography

Major publications by the team in recent years

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] D. LAVENIER, M. GIRAUD. *Bioinformatics Applications*, in "Reconfigurable Computing: Accelerating Computation with Field-Programmable Gate Arrays", M. B. GOKHALE, P. S. GRAHAM (editors), Springer, 2005, http://dx.doi.org/10.1007/0-387-26106-0_8.

[8] 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/.

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

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

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] A. LIEFOOGHE. *Matrices score-position, algorithmes et propriétés*, Ph. D. Thesis, Université de Lille 1, 2008.

[12] J.-S. VARRÉ. *Algorithmes pour la comparaison de génomes et la recherche de signaux cis-régulateurs*, Habilitation à Diriger des Recherches, Université de Lille 1, 2008.

Articles in International Peer-Reviewed Journal

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

[14] A. CHENUIL, E. EGEA, C. ROCHER, H. TOUZET, J.-P. FÉRAL. *Does hybridization increase evolutionary rate ? Data from the 28S-rDNA D8 domain in echinoderms*, in "Journal of Molecular Evolution", vol. 67, n^{o} 5, 2008, p. 539-550, http://www.springerlink.com/content/526u861u76128q27/fulltext.html.

[15] 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.

[16] A. FONTAINE, H. TOUZET. *Computational identification of protein-coding sequences by comparative analysis*, in "International Journal of Data Mining and Bioinformatics", to appear, 2009.

[17] 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.

[18] 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", to appear, 2008.

International Peer-Reviewed Conference/Proceedings

[19] J. BYRKA, S. GUILLEMOT, J. JANSSON. *New Results on Optimizing Rooted Triplets Consistency*, in "Proceedings of the 19th International Symposium on Algorithms and Computation (ISAAC 2008)", Lecture Notes in Computer Science, vol. 5369, 2008, p. 484–495, http://www.springerlink.com/content/165qk27gp176x21t/.

[20] M. GIRAUD. *Not so many runs in strings*, in "Int. Conf. on Language and Automata Theory and Applications (LATA 08)", Lecture Notes in Computer Science (LNCS), vol. 5196, 2008, p. 232–239, http://www.lifl.fr/~giraud/publis/giraud-lata-08.pdf.

[21] S. GUILLEMOT. *FPT algorithms for path-transversals and cycle-transversals problems in graphs*, in "Proceedings of Parameterized and Exact Computation: Third International Workshop (IWPEC 2008)", Lecture Notes in Computer Science, vol. 5018, 2008, p. 129–140, http://www.springerlink.com/content/1621103545417h7n/.

[22] S. GUILLEMOT. *Parameterized complexity and approximability of the SLCS problem*, in "Proceedings of Parameterized and Exact Computation: Third International Workshop (IWPEC 2008)", Lecture Notes in Computer Science, vol. 5018, 2008, p. 115–128, http://www.springerlink.com/content/qqm6mp6hn6np4147/.

[23] R. KOLPAKOV, G. KUCHEROV. *Searching for gapped palindromes*, in "Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching (CPM), June 18-20, 2008, Pisa (Italy)", Lecture Notes in Computer Science, vol. 5029, Springer Verlag, 2008, p. 18–30, http://www.springerlink.com/content/v458251304h45533/.

[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", to appear, 2009.

[25] 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.springerlink.com/content/2280v0131631528r/.

[26] 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/.

National Peer-Reviewed Conference/Proceedings

[27] J. ALLALI, Y. D'AUBENTON-CARAFA, C. CHAUVE, A. DENISE, C. DREVET, P. FERRARO, D. GAUTHERET, C. HERRBACH, F. LECLERC, A. DE MONTE, A. OUANGRAOUA, M.-F. SAGOT, C. SAULE, M. TERMIER, C. THERMES, H. TOUZET. *Benchmarking RNA secondary structure comparison algorithms*, in "Proceedings of the 9th Open Days in Biology, Computer Science and Mathematics (JOBIM), June 30-July 2, 2008, Lille (France)", (short talk with poster), 2008, p. 67–68, http://www.labri.fr/publications/mabiovis/2008/ACDDFGHLDOSSTTT08/brasero_bench1_.pdf.

[28] M. GÎRDEA, G. KUCHEROV, L. NOÉ. *Protein sequence alignment via anti translation*, in "Proceedings of the 9th Open Days in Biology, Computer Science and Mathematics (JOBIM), June 30-July 2, 2008, Lille (France)", (short talk with poster), 2008, p. 157–158, http://www.lifl.fr/~girdea/at/files/jobim044.pdf.

[29] J.-É. SYMPHOR, A. MANCHERON, L. VINCESLAS, P. PONCELET. *Le* FIA *: un nouvel automate permettant l'extraction efficace d'itemsets fréquents dans les flots de données*, in "Proceedings of the 8th Extraction et Gestion des Connaissances (egc)", Revue des Nouvelles Technologies de l'Information, vol. RNTI-E-11 (1), Cépaduès-Éditions, 2008, p. 157–168, http://www.lgi2p.ema.fr/~poncelet/publications/papers/SymphorEGC087.pdf.

[30] L. VINCESLAS, J.-É. SYMPHOR, A. MANCHERON, P. PONCELET. *FIASCO : un nouvel algorithme d'extraction d'itemsets fréquents dans les flots de données*, in "Proceedings of the 8th Extraction et Gestion des Connaissances (EGC)", Revue des Nouvelles Technologies de l'Information, vol. RNTI-E-11 (1), Cépaduès-Éditions, 2008, p. 235–236.

Scientific Books (or Scientific Book chapters)

[31] J. JANSSON. *Phylogenetic Tree Construction from a Distance Matrix – 1989; Hein*, in "Encyclopedia of Algorithms", M.-Y. KAO (editor), Springer, 2008, p. 651-653, http://www.springerlink.com/content/n6002jk482656404/.

[32] J. JANSSON. *Perfect Phylogeny (Bounded Number of States) – 1997; Kannan, Warnow*, in "Encyclopedia of Algorithms", M.-Y. KAO (editor), Springer, 2008, p. 644-647, http://www.springerlink.com/content/p712322k476p6m0t/.

[33] J. JANSSON. *Directed Perfect Phylogeny (Binary Characters) – 1991; Gusfield*, in "Encyclopedia of Algorithms", M.-Y. KAO (editor), Springer, 2008, p. 246-248, http://www.springerlink.com/content/qx64143355344m57/.

[34] J. JANSSON, W.-K. SUNG. *The Maximum Agreement of Two Nested Phylogenetic Networks*, in "New Topics in Theoretical Computer Science", NOVA Publishers, 2008, p. 119-141, http://www.df.lth.se/~jj/Publications/nested11.pdf.

[35] G. KUCHEROV, D. SOKOL. *Approximate Tandem Repeats*, in "Encyclopedia of Algorithms", M.-Y. KAO (editor), Springer, 2008, p. 48–51, http://www.springerlink.com/content/r7376614263rv3h0/.

Other Publications

[36] S. CABOCHE, M. PUPIN, V. LECLÈRE, P. JACQUES, G. KUCHEROV. *NORINE: a platform dedicated to nonribosomal peptides*, 9th Open Days in Biology, Computer Science and Mathematics (JOBIM), June 30-July 2, 2008, Lille (France), 2008.

[37] S. CABOCHE, M. PUPIN, V. LECLÈRE, P. JACQUES, G. KUCHEROV. *NORINE: a public resource for nonribosomal peptides*, EMBnet 2008 - 20th Anniversary Conference, 2008, http://issuu.com/lfalquet/docs/embnet_news_14_3/64.

[38] S. CABOCHE, M. PUPIN, V. LECLÈRE, P. JACQUES, G. KUCHEROV. *NORINE: database and efficient algorithms dedicated to nonribosomal peptides*, ECCB08 European Conference on Computational Biology, 2008, http://www.eccb08.org/themes/default/poster/poster/posters-B.html#8.

[39] A. DARRACQ, J.-S. VARRÉ, A. COURSEAUX, L. MARÉCHAL-DROUARD, P. TOUZET. *Evolution of the mitochondrial genome in beet. A comparative genomic study at the intra-specific level*, XX International Congress of Genetics, 2008.

References in notes

[40] S. ALTSCHUL, Y. MADDEN, A. SCHÄFFER, J. ZHANG, Z. ZHANG, W. MILLER, D. LIPMAN. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, in "Nucleic Acids Research", vol. 25, 1997, p. 3389-3402.

[41] M. ANSARI, G. YADAV, R. GOKHALE, D. MOHANTY. *NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases*, in "Nucleic Acids Res.", vol. 32(Web Server issue), 2004, p. W405-W413.

[42] 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/.

[43] D. BROWN. *Optimizing Multiple Seeds for Protein Homology Search*, in "IEEE Transactions on Computational Biology and Bioinformatics (IEEE TCBB)", vol. 2, n^{o} 1, january 2005, p. 29–38.

[44] M. CSÜRÖS, B. MA. *Rapid homology search with neighbor seeds*, in "Algorithmica", vol. 48, n^{o} 2, june 2007, p. 187–202.

[45] S. R. EDDY. *A Model of the Statistical Power of Comparative Genome Sequence Analysis*, in "PLoS Biology", vol. 3(1), 2005.

[46] R. ELKON, C. LINHART, R. SHARAN, R. SHAMIR, Y. SHILOAH. *Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells.*, in "Genome Res", vol. 13, n^{o} 5, 2003, p. 773-80.

[47] M. FARACH-COLTON, G. M. LANDAU, C. SAHINALP, D. TSUR. *Optimal spaced seeds for faster approximate string matching*, in "Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP'05), Lisboa (Portugal)", Lecture Notes in Computer Science, vol. 3580, Springer-Verlag, 2005, p. 1251–1262.

[48] S. FENG, E. TILLIER. *A fast and flexible approach to oligonucleotide probe design for genomes and gene families*, in "Bioinformatics", vol. 23, n^{o} 10, 2007, p. 1195–1202.

[49] A. FONTAINE, M. GIRAUD, L. NOÉ, H. TOUZET. *Graines espacées et recherche d'ARN non-codants*, in "Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)", (poster), 2007.

[50] E. K. FREYHULT, J. P. BOLLBACK, P. P. GARDNER. *Exploring genomic dark matter: a critical assessment of the performance of homology search methods on non-coding RNA*, in "Genome Research", 2006.

[51] E. K. FREYHULT, J. P. BOLLBACK, P. P. GARDNER. *Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA.*, in "Genome Res", vol. 17, n^{o} 1, 2007, p. 117–125.

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

[53] S. GRIFFITHS-JONES, A. BATEMAN, M. MARSHALL, A. KHANNA, S. R. EDDY. *RFAM: an RNA family database*, in "Nucleic Acids Research", vol. 31, n^{o} 1, 2003, p. 439-441, http://rfam.janelia.org/.

[54] S. HO SUI, J. MORTIMER, D. ARENILLAS, J. BRUMM, C. WALSH, B. KENNEDY, W. WASSERMAN. *oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.*, in "Nucleic Acids Res", vol. 33, n^{o} 10, 2005, p. 3154-64.

[55] L. ILIE, S. ILIE. *Fast computation of good multiple spaced seeds*, in "Proceedings of the 7th International Workshop in Algorithms in Bioinformatics (WABI), Philadelphia (USA)", Lecture Notes in Bioinformatics, vol. 4645, Springer-Verlag, Sept. 2007, p. 346–358.

[56] R. KOLPAKOV, G. BANA, G. KUCHEROV. *mreps: efficient and flexible detection of tandem repeats in DNA*, in "Nucleic Acid Research", accepted for publication for the special issue on Web software, vol. 31, n^{o} 13, 2003, p. 3672-3678.

[57] R. KOLPAKOV, G. KUCHEROV. *Maximal Repetitions in Words or How to Find all Squares in Linear Time*, Technical report, n^{o} 98-R-227, LORIA, 1998.

[58] H. KONG. *Generalized Correlation Functions and Their Applications in Selection of Optimal Multiple Spaced Seeds for Homology Search*, in "Journal of Computational Biology", vol. 14, n^{o} 2, Mar. 2007, p. 238–254.

[59] D. KONZ, M. MARAHIEL. *How do peptide synthetases generate structural diversity?*, in "Chemistry & Biology", vol. 6 (2), 1999, p. R39-R48.

[60] 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° 2, 2006, p. 553–569, http://www.worldscinet.com/jbcb/04/0402/S0219720006001977.html.

[61] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *Subset Seed Automaton*, in "Proceedings of the 12th International Conference on Implementation and Application of Automata (CIAA), Prague (Czech Republic), July 16-18, 2007", J. HOLUB, J. ZDAREK (editors), Lecture Notes in Computer Science, vol. 4783, Springer Verlag, 2007, p. 180–191, http://www.springerlink.com/content/y824l20554002756/.

[62] T. LI, K. FAN, J. WANG, W. WANG. *Reduction of Protein Sequence Complexity by Residue Grouping*, in "Journal of Protein Engineering", vol. 16, 2003, p. 323–330.

[63] B. MA, J. TROMP, M. LI. *PatternHunter: faster and more sensitive homology search*, in "Bioinformatics", vol. 18, n° 3, March 2002, p. 440–445.

[64] D. MAK, G. BENSON. *All hits all the time: parameter free calculation of seed sensitivity*, in "Proceedings of the 5th Asia Pacific Bioinformatics Conference (APBC)", 2007, p. 317–326.

[65] H. MOOTZ, D. SCHWARZER, M. MARAHIEL. *Ways of assembling complex natural products on modular nonribosomal peptide synthetases*, in "ChemBioChem", vol. 3(6), 2002, p. 490-504.

[66] L. MURPHY, A. WALLQVIST, R. LEVY. *Simplified amino acid alphabets for protein fold recognition and implications for folding*, in "Journal of Protein Engineering", vol. 13, 2000, p. 149–152.

[67] N. NAAMANE, J. VAN HELDEN, D. L. EIZIRIK. *In silico identification of NF-kappaB-regulated genes in pancreatic beta-cells*, in "BMC Bioinformatics", vol. 8(55), 2007.

[68] F. NICOLAS, E. RIVALS. *Hardness of Optimal Spaced Seed Design*, in "Proceedings of the 16th Annual Symposium on Combinatorial Pattern Matching (CPM), Jeju Island (Korea)", A. APOSTOLICO, M. CROCHEMORE, K. PARK (editors), Lecture Notes in Computer Science, vol. 3537, Springer-Verlag, 2005, p. 144–155.

[69] L. NOÉ, G. KUCHEROV. *Improved hit criteria for DNA local alignment*, in "BMC Bioinformatics", vol. 5, n° 149, 2004.

[70] O. PERRIQUET, H. TOUZET, M. DAUCHET. *Finding the common structure shared by two homologous RNAs*, in "Bioinformatics", vol. 19, 2003, p. 108-116, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12499300&dopt=Abstract.

[71] 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.

[72] P. PETERLONGO, N. PISANTI, F. BOYER, M.-F. SAGOT. *Lossless Filter for Finding Long Multiple Approximate Repetitions Using a New Data Structure, the Bi-factor Array*, in "SPIRE", 2005, p. 179–190.

[73] E. RIVAS, R. J. KLEIN, T. A. JONES, S. R. EDDY. *Computational identification of noncoding RNAs in E. coli by comparative genomics*, in "Current Biology", vol. 11, 2001, p. 1369-1373.

[74] M. TOMPA, N. LI, T. L. BAILEY, G. M. CHURCH, B. D. MOOR, E. ESKIN, A. V. FAVOROV, M. C. FRITH, Y. FU, W. J. KENT, V. J. MAKEEV, A. A. MIRONOV, W. S. NOBLE, G. PAVESI, G. PESOLE, M. REGNIER, N. SIMONIS, S. SINHA, G. THIJS, J. VAN HELDEN, M. VANDENBOGAERT, Z. WENG, C. WORKMAN, C. YE, Z. ZHU. *Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites*, in "Nature Biotechnology", vol. 23, n^{o} 1, 2005, p. 137 - 144.

[75] H. TOUZET, O. PERRIQUET. *CARNAC: folding families of related RNAs*, in "Nucleic Acids Research", vol. 32 (Supplement 2), 2004, p. 142-145, http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_2/W142.

[76] H. TOUZET. *Tree edit distance with gaps*, in "Information Processing Letters", vol. 85, n^{o} 3, 2003, p. 123-129.

[77] H. TOUZET. *A linear tree edit distance algorithm for similar ordered trees*, in "Proc. of the 16th Annual Symposium Combinatorial Pattern Matching (CPM 2005), Jeju Island, Korea, June 19-22, 2005", Lecture Notes in Computer Science, vol. 3537, Springer Verlag, 2005, p. 334-345.

[78] H. TOUZET. *Comparative analysis of RNA genes: the CaRNAc software*, N. BERGMAN (editor), vol. Methods in Molecular Biology, Special issue on comparative genomics I, Humana Press, 2007, p. 465-473.

[79] C. TUGGLE, Y. WANG, O. COUTURE, L. QU, J. UTHE, D. KUHAR, J. LUNNEY, D. NETTLETON, J. DEKKERS, M. BEARSON S. *Characterizing the porcine transcriptional regulatory response to infection by Salmonella: identifying putative new NFkB direct targets through comparative bioinformatics*, 2007, http://eadgene.org/.

[80] H. WANG, C. PERNG, W. FAN, S. PARK, P. YU. *Indexing weighted sequences in large databases*, in "ICDE", 2003, http://citeseer.ist.psu.edu/wang03indexing.html.

[81] C. WOELK, F. OTTONES, C. PLOTKIN, P. DU, C. ROYER, S. ROUGHT, J. LOZACH, R. SASIK, R. KORNBLUTH, D. RICHMAN, J. CORBEIL. *Interferon Gene Expression following HIV Type 1 Infection of Monocyte-Derived Macrophages*, in "AIDS Res Hum Retroviruses", vol. 20(11), 2004, p. 1210-22.

[82] L. ZHANG. *Superiority of Spaced Seeds for Homology Search*, in "IEEE Transactions on Computational Biology and Bioinformatics (IEEE TCBB)", vol. 4, n^{o} 3, 2007, p. 496–505.

[83] L. ZHOU, L. FLOREA. *Designing sensitive and specific spaced seeds for cross-species mRNA-to-genome alignment*, in "Journal of Computational Biology", vol. 14, n^{o} 2, Mar. 2007, p. 113–130.