INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team sequoia

Algorithms for large-scale sequence analysis

Futurs

THEME BIO

Table of contents

1. Team

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

Head of the team
Gregory Kucherov [DR CNRS, HdR]

Administrative assistant
Axelle Magnier [INRIA]

Full-time researchers
Hélène Touzet [CR CNRS, HdR]
Mathieu Giraud [CR CNRS, from October 2006]

Faculty members
Maude Pupin [MC, Université Lille 1]
Jean-Stéphane Varré [MC, Université Lille 1]
Laurent Noé [MC, Université Lille 1]

PhD students
Mathieu Defrance [MESR fellowship, from October 2003]
Aude Liefooghe [MESR fellowship, from October 2004]
Arnaud Fontaine [MESR fellowship, from October 2005]
Ségolène Caboche [INRIA/Region fellowship, from October 2006]

Associate member
Max Dauchet [Pr, Université Lille 1, HdR]

Visiting scientists
Liviu Ciortuz [University of Iasi, Romania, June 1st – July 31]
Anna Gambin [Warsaw University, July 30 – August 19]
Slawomir Lasota [Warsaw University, July 30 – August 19]
Mikhail Roytberg [Russian Academy of Sciences, October 30 – December 22]

Internships
Julien Ferté [ENS Cachan, 26/06/2006 – 25/08/2006]
Ewa Makosa [Warsaw University, 10/07/2006 – 25/08/2006]
Oana Ratoi [University of Iasi, Romania, 01/02/2006 – 31/07/2006]
Sébastien Tonon [Université Lille 1, 01/04/2006 – 31/07/2006]
Djamel Zitouni [Université Lille 1, 03/01/2006 – 30/06/2006]

2. Overall Objectives

2.1. Overall Objectives

Keywords: *RNA structures*, *algorithmics*, *bioinformatics*, *computational biology*, *discrete algorithms*, *genomic sequences*, *protein sequences*, *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 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 its central component.

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 large-scale sequence comparison, prediction, analysis and manipulation of RNA secondary structures, identification and processing of regulatory sequences. 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 alternative problems for 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. 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 in 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 [22], is by far the most widely used bioinformatics software. Since 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 this sequences bear a biological function; moreover, similar sequences are likely to correspond to similar biological functions and/or to a common evolutionary ancestor.

3.1.1. Spaced-seed-based similarity search

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 [38]. 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 last years, see [25], [44], [37] 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 [41] [6] developed by group members (see Section 4.2).

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 3.5.3) 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[4], [30], [40].

3.1.2. 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[3]. Developed algorithmic techniques have been implemented in the `mreps` software [35] (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. [42]) that is important in numerous applications that will be mentioned later on in this report.

3.1.3. Spaced seeds for protein alignment

Spaced seeds (see Section 3.1.1) 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 ([24] 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 [12], 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. The main one is defining letters of seed alphabet, that is corresponding subsets of pairs of amino acids. The choice of those letters is crucial for constructing seeds with good selectivity/sensitivity ratio. On the other hand, it is also very important for the efficient implementation of the search procedure: certain seeds, namely those that induce an equivalence relation on pairs of sequence keys, allow an implementation by direct hashing and are therefore advantageous. 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 be bring an important improvement to the whole technology.

3.2. Non-coding RNA analysis

Keywords: *RNA*, *base pairings*, *secondary structure*, *structure alignment*, *structure inference*.

As mentioned in the introduction to this section, 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*, has 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 modelled 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, together with the sequencing of full genomes and the availability of an increasing number of families of homologous RNA genes. There is currently a need for computational tools for a systematic analysis of those genes, analogous to those available for protein-coding genes.

3.2.1. 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. The idea is that the structure is preserved by evolution, and mutations observed between homologous RNA sequences should not be distributed 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 reduces partially to the problem of determining if sequences actually share a common structure. We developed recently a CARNAC software for structure prediction [7], [8] (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.

3.2.2. 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 3.1), 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 [1], [46], [47]. 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 [26]. 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.

3.3. 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. Modelling, identification and analysis of TFBSs is one of major bioinformatics challenges.

3.3.1. 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 [45].

The most popular model of TFBSs is given by *Position Weight Matrices* (PWMs), which are probalilistic 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 modelled by PWMs [28], [33]. 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 [27]. This research direction is therefore very promising and has still a lot of progress to be made.

3.3.2. 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 [48] and would lead to an advanced TFBS retrieval system.

3.4. Non-ribosomal peptide synthesis

Keywords: *amino acids*, *non-ribosomal 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, it 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 [36], [39].

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-conventionnal 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 [23] 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 is to design a comprehensive computational tool for working with non-ribosomal peptides. Such as tool should include several components. First, it should include a complete database of annotated NRPS peptides. The first prototype of such a database, called NORINE, has already been implemented and will be described in Section 4.5. Second, the tool should 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 6.1).

3.5. General models and tools

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

In contrast to Sections 3.1-3.4, 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 3.1-3.4.

3.5.1. Discrete algorithms

3.5.1.1. 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 modelled 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.

3.5.1.2. 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 3.1) use an index for storing seed keys. Indexing problems appear in RNA matching (as mentioned in Section 3.2) as well as in PWM search (Section 3.3). Thus, designing efficient index structures is crucial for many of our research topics and holds therefore a particuliar place within the scope of our studies. Note that we participate in a collaboration on efficient index structures within an INRIA ARC project led by the SYMBIOSE team (see Section 6.2).

3.5.2. Statistics and discrete probability

This area is of more applied nature for our team but still plays an important role in our research work. Our approach here is generally not to develop original computational techniques but rather to be "active users" of existing statistical and probabilistic methods.

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.

3.5.3. 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 are 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 3.2).

Another way to enhance computing performances is to use *specialized computer architectures* to obtain a fine-grained parallelism [5]. We collaborate with the SYMBIOSE project-team (INRIA-Rennes) that builds prototypes designed to index large amounts of data (see Section 6.2). We also plan to further pursue this line of research by considering a *Genome on Chip* architectural paradigm. The main goal of those projects is to

index complete genomes to allow fast queries of different types, ranging from sequence similarities queries to structure-based queries (approximate RNA pattern matching, see Section 3.2).

4. Software

4.1. Introduction

Software development is an important part of our work as many of the algorithmic techniques we develop are implementated in experimental or deliverable software. We maintain a server accessible via http://bioinfo.lifl.fr/ for distributing our software and executing it through web interfaces. Our main software programs are also available through the *Génopole* website[1]. Below we first present software programs that are currently actively developed in the team.

4.2. YASS suite

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

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

YASS [41] [6] 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 3.1), and the other is a new *hit criterion* that specifies the way that the seeds are used to detect potential similarity regions. Besides the Web-server of our team, version 1.11 of YASS is available from the INRIA software web page[2].

HEDERA is an accompanying program for designing spaced seeds and transition-constrained seeds, created to design new seeds for the YASS software. HEDERA is available from the YASS Web page accompanied with a user documentation.

4.3. caRNAc suite

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

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

On the subject of RNA analysis, the CARNAC program for RNA structure prediction is currently made available to the community. 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 [31], and will be presented in a book chapter devoted to comparative genomics [14]. The current release includes a home-made Java applet – RNAfamily – that is devoted to the visualization of homologous RNA structures, as well the Naview 2D viewer. In future, the CARNAC suite should be extended to incorporate upcoming results in structure comparison (pairwise and multiple) and gene prediction.

4.4. TFM suite

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

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

[1] http://www.genopole-lille.fr
[2] http://www.inria.fr/valorisation/logiciels/vie.fr.html

Our research on cis-regulatory regions described in Section 3.3 is being implemented in a platform devoted to the location and processing of Position Weight Matrices. An embryo of this platform already exists in the TFM-EXPLORER software, dedicated to the inference of locally over-represented motifs in mammalian genomes [10]. The server includes pre-computed background models for Human, Mouse and Rat genomes derived from annotated genes with REFSEQ identifiers [43] available from the UCSC Genome Browser assembly [34] (release hg18, mm8, rn3). Promoter regions corresponding to 10 000 bp upstream and 1000 bp downstream Transcription Start Sites are used to build background models. Potential TFBSs are exhaustively pre-computed for all TRANSFAC and JASPAR vertebrates matrices. TFM-Explorer is accompanied by the TFM-Scan program [21], that implements the methods that we have developed to speed up the location of PWM matrices on a sequence (see Section 5.3.2).

4.5. Norine

Keywords: *database*, *non-ribosomal peptide synthesis*.

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

We develop a database of NRPS peptides called NORINE[3]. The list of NRPS peptides is obtained from scratch as there is no centralized resource of these data. Among existing resources, NRPS-PKS[4] contains only 20 peptides and is focused on the synthases, other resources (PubChem[5] or ChEBI[6]) contain some of NRPS peptides and many other small biological molecules. Similarly, in the literature, there is no complete review devoted specifically to NRPS peptides. Therefore, we had to explore publications appeared since the 70's to compile an exhaustive list of known NRPS peptides. Today NORINE contains about 700 peptides (of which 328 are currently available from the web site), described in about 350 publications. The entries contain various annotations of those 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. One can query the annotations via a web interface to select the NRPS peptides that correspond to a search criteria.

4.6. Other software

Several software programs have previously been developed by group members and are currently used, maintained and distributed from our software server or through other means.

- `mreps` (http://bioinfo.lifl.fr/mreps, see Section 3.1), is a program that enables one to compute *all* tandem repeats in a DNA sequence (without any restriction on the size of the repeated unit) by a single run of the program that takes several seconds on a sequence of several megabases (typical size of a bacterial genome). The core of the `mreps` method is constituted by a very efficient algorithm that computes all so-called *maximal repetitions*.

  `mreps` can be queried through its Web page[7], as well as through the BIOWEB server of the Pasteur Institute[8] and the *Tandem Repeat Data Base (TRDB)*[9]. It is distribuited from the INRIA free software server[10].

- `grappe` (http://www.inria.fr/valorisation/logiciels) is a program that simultaneously searches in a text for several patterns, each of them composed of a list of fragments (words) separated by "jokers" (don't care symbols) of bounded or non-bounded length. A special version of `grappe` for processing DNA/RNA sequences that has been used in our work on regulatory sequence analysis (see Section 3.3).

[3] **non**-**ri**bosomal peptides, with **ine** as a typical ending of names of non-ribosomal peptides
[4] http://www.nii.res.in/nrps-pks.html
[5] http://pubchem.ncbi.nlm.nih.gov
[6] http://www.ebi.ac.uk/chebi
[7] http://bioinfo.lifl.fr/mreps/
[8] http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html
[9] http://tandem.bu.edu/trf/trf.html
[10] http://www.inria.fr/valorisation/logiciels

- HUGO (http://bioinfo.lifl.fr/HUGO, *Hierarchical Union of Genes from Operons*) is a program that detects conserved clusters of genes among several procaryotic species. It infers how genome rearrangements affect genome organization, and more precisely clusters of genes (sets of co-located genes). The input of HUGO is a list of species, each described as a set of operons, i.e. ordered lists of (possibly duplicated) genes. Out of this, HUGO computes a set of super-operons, where a super-operon is a set of genes made of the union of conserved and similar operons. A particularity of HUGO is that the output is presented as a clusterisation with associated probability for each node of the clusterisation. The core of the HUGO algorithm is based on graph-theoretic techniques.

5. New Results

5.1. Sequence similarity and repetitions

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

5.1.1. Estimation of seed sensitivity

A journal version of paper [12] appeared this year. The paper presents a general approach to automatically obtain an efficient algorithm for various instances of the seed sensitivity problem. The approach treats separately three components of the seed sensitivity problem – a set of target alignments, an associated probability distribution, and a seed model – that are specified by distinct finite automata. We showed that once these three components are specified, one can construct, using a single general method, a dynamic programming algorithm for computing seed sensitivity.

The proposed approach has then been applied to a new seed model, called *subset seed* and an efficient automaton construction for the set of alignments detected by subset seeds has been presented. This automaton and the whole associated algorithm has been implemented in the HEDERA software (see Section 4.2).

5.1.2. Seeds for protein search

The formalism of subset seeds, mentioned in the previous paragraph, allows to take into account, in a subtle way, different degrees of affinity between pairs of letters of the sequence alphabet. With this motivation in mind, we studied the problem of similarity search in protein sequences using the subset seeds paradigm. This work was intensified during the summer stay of our polish colleagues within the ECO-NET cooperation (see Section 6.3.2): a 1.5-month internship of Ewa Makosa, a master student from Warsaw University, as well as a stays of A. Gambin and S. Lasota.

As a result, we succeeded to overcome the main difficulty of this approach, as we proposed a method to design efficient seed alphabets. Based on these alphabets, we were able to design efficient seeds according to the technique developed for the DNA case. Preliminary experiments show that this approach allows us to obtain a selectivity/sensitivity ratio comparable to (or even, in certain cases, better than) that of BLAST. These results are interesting as the formalism of subset seeds is weaker and less costly than the method of BLAST. Currently this research direction is continued in collaboration with M. Roytberg, another partner of the ECO-NET project. A paper describing these studies is under preparation.

5.1.3. Statistics of genomic word counts

In collaboration with Prof. Miklós Csűrös from the University of Montréal, we studied the distribution of oligonucleotide counts in genomic sequences. As mentioned in Section 3.5.2, functional elements in a genome sequence can be computationally identified only with respect to an adequate statistical model of non-functional DNA sequences (*null model*). A sequence feature can be conjectured to have a functional role if it is observed too often or too rarely in the genome with respect to the expected frequency defined by the null model. The validity of such inference depends on the precise characterization of feature occurrences in neutrally evolving DNA. We proposed that the distribution of DNA words in genomic sequences is primarily characterized by a double Pareto-lognormal distribution, which explains lognormal and power-law features found across all known genomes. Such a distribution may be the result of random evolution by a copying process, and is therefore useful in characterizing sequence features evolving without functional pressure. A paper describing this study is submitted to an international journal.

5.1.4. Layout problems for interval graphs

Interval graphs are extensively used in bioinformatics, typically to model the genome physical mapping problem, which is the problem of reconstructing the relative positions of DNA fragments, called *clones*, out of information of their pairwise overlaps. However, interval graphs appear also in other situations in bioinformatics, such as for gene structure prediction for example. In [29], interval graphs are used to model temporal relations in protein-protein interactions. In that paper, an optimal linear arrangement (OLA) of an interval graph models an "optimal" molecular pathway, and the problem of efficiently computing this arrangement is explicitly raised.

With this motivation, we studied in paper [18] the OLA problem on interval graphs. Several linear layout problems that are NP-hard on general graphs are solvable in polynomial time on interval graphs. We proved that, quite surprisingly, optimal linear arrangement of interval graphs is NP-hard. The same result holds for permutation graphs. We presented a lower bound and a simple and fast 2-approximation algorithm based on any interval model of the input graph. This is a joint work with J. Cohen (Loria, Nancy), D. Kratsch (University of Metz) and F. Fomin and P. Heggernes from the University of Bergen (Norway).

5.2. RNA genes and RNA structures

Keywords: *RNA*, *base pairings*, *secondary structure*, *structure alignment*, *structure inference*.

5.2.1. RNA structure comparison

In the scope on RNA comparison, we have addressed the problem of comparing similar RNA sequences with short evolutionary distance. In presence of a family of homologous RNAs, the number of errors can be bounded in advance by a finite parameter. In this context, we have shown that it is likely to speed up the computation process by carefully pruning the computation space. We have proposed a linear-time algorithm for the problem, which is as far as we know the fastest algorithm existing for the tree comparison problem. A journal version of this work appeared this year [15]. The algorithm has been implementated by Djamel Zitouni during his master internship.

We also obtained new results concerning the comparison of RNA structures encoded by arc-annotated sequences. Arc-annotated sequences are the most expressive combinatorial representation to model RNA evolution. We have defined a unifying framework, which we called the *alignment hierarchy* [16]. We have shown that the alignment hierarchy encompasses main existing models. This study is relevant from both practical and theoretical viewpoint. We have provided two polynomial time algorithms to compare arc-annotated sequences of nested type with arc-altering and arc-breaking operations, whereas when considering other models, the problem is NP-hard. We also proved a new NP-completeness result, that enhances understanding of the complexity of arc-annotated sequences comparison. This result sheds a new light on the border between tractability and untractability when dealing with arc-annotated sequences. Ongoing work is concerned with the implementation of the two polynomial algorithms, enriched with an evolutionnary model taking into account affine gap weights, constraints coming from the primary structure, and local search.

5.2.2. RNA gene prediction

We studied a classification procedure for coding and non-coding genes based on evolutionary patterns of DNA sequences. 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. On the other hand, RNA genes tend to support compensatory mutations that preserve the formation of the base pairings involved in the structure of the molecule. This observation gave rise to the definition of two statistical models. A protein coding model uses a graph-theoretic encoding of all the six possible reading frames of each sequence. An RNA non-coding model is based on the caRNAc software (described in Section 4.3). We performed a large-scale validation on two biological databases (RFAM for non-coding genes [32] and Pandit and coding genes [49]), as well as on random data. On non-coding RNAs, this research direction is carried out in collaboration with L. Ciortuz , and a paper describing the studies of the coding model is under preparation.

5.3. Cis-regulatory sequence analysis

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

5.3.1. PWM matching problem

We proposed an efficient algorithm for the PWM matching problem in presence of a large set of PWMs [21]. The foundation of the method is to pre-process PWM matrices and to store scores in a multi-index table. The index is optimized with respect to the set of matrices, the P-value threshold for score cutoff and an amount of memory. Hence, the index can be built in advance and stored into the main memory giving rise to a very efficient score computation for all matrices on a given sequence. This algorithm is eight times faster than the brute-force algorithm. We also investigated the problem of PWM matching for similar matrices. In this perspective, we formulated exact relationships between the set of occurrences of PWMs, that allow to estimate the redundancy of the occurrences. We believe that these results are of more general interest, and may be used in larger contexts for assessing the significativity of multiple occurrences. This question arises frequently when studying regulatory sequences and putative transription factor binding sites. Another virtue of this analysis is that it helps to cope with redundant site occurrences, which is a usual problem when one works with public databases.

5.3.2. Mining PWM predictions

Besides the brute identification of TFBSs modeled by PWMs, we presented a complementary method that searches for locally overrepresented PWM sites in a set of coregulated genes [10]. The algorithm, which we have named TFM-Explorer, associates motif overrepresentation with comparative genomics, allowing for multiple species to be included. One novel feature of the method is that it takes advantage of the spatial conservation of cis-regulatory elements, when it exists. More precisely, TFM-Explorer relies on three main principles. The first is that the background distribution used to assess the statistical significance of overrepresented motifs is a local model that depends on the location on the sequence with respect to the TSS. This allows us to cope with large heterogeneous regulatory regions, including proximal cis-regulatory elements as well as distal enhancers. Second, it is possible to combine background models between sequences, which makes the method capable to cope with multiple species. In contrast with other phylogenetic footprinting approaches, genes do not need to be orthologous, and conserved TFBSs are not expected to be surrounded by similar regions that can be easily aligned. Lastly, we use spatial conservation as supplementary information, for which we have developed an algorithm that is able to identify the portion of sequences with local overrepresentation without prior knowledge of either the size or the location of the involved region. This allows us to infer short regions exhibiting a local signal, as well as large regions when we have to identify cis-regulatory motifs that show no spatial conservation.

5.4. Non-ribosomal peptide synthesis

Keywords: *amino acids*, *non-ribosomal peptide synthesis*, *synthetase*.

As presented in Section 3.4, there does not exist today a computer tool that would allow one to manipulate (retrieve, compare, search, ...) numerous peptides issued from the non-ribosomal synthesis pathway. Note that the number of known such peptides is counted by hundreds and is still growing. On the other hand, no review article or web resource features a complete list of such peptides. 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 aminoacids, non-ribosomal peptides contain more than 400 different monomers. Finally, they have several interesting activities, such as antibiotic, anti-inflammatory, antithrombotic, antitumor, calmodulin antagonist, immunomodulating, protease inhibitor, siderophore, surfactant, and toxin.

The first goal of this project was then to create a database containing a possibly complete list of annotated non-ribosomal peptides. This work started this year, within the Master diploma work of Ségolène Caboche, and resulted in the NORINE prototype, described in Section 4.5. First presentations of this work have been made in a short talk to the JOBIM conference this year [17], and in a poster at the conference of the Royal Society of Chemistry held this year in Cambridge, UK [20]. A submission is currently being prepared to the journal *Natural Product Reports*.

On the other hand, we studied algorithms of comparing NRPS molecules, represented as non-oriented labelled graphs. As a result, an efficient algorithm for this task has been developed and implemented and will be incorporated into the NORINE system.

6. Other Grants and Activities

6.1. 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[11].
- Research on *cis-regulatory region analysis* relies on a collaboration with UMR 8161 (Biological Institute of Lille, CNRS – Lille Pasteur Institut– University Lille 1 – University of Lille 2, Pr. Delaunoy), and more particularly with the group led by professor C. Abbadie. This research theme also benefits from regular relationships with UMR 8576 (Structural and Functional Glycobiology, CNRS – University Lille 1, Pr. Michalski) and UMR 8090 (Genetics of Multifactorial Diseases, CNRS – Lille Pasteur Institute, Pr. Froguel – University Lille 2).
- The project on *non-ribosomal peptide synthesis* stems from a collaboration with the laboratory ProBioGem (*Laboratoire des Procédés Biologiques Génie Enzymatique et Microbien*), headed by Pr. Guillochon, University Lille 1. This laboratory develops methods to produce and extract active peptides in agriculture or food. A co-supervised PhD student (Ségolène Caboche) started her PhD work on this subject in October 2006.
- 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 modelled 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 – FRE CNRS, headed by Prof. Vandenbunder, and then by Prof. Blossey). 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 regional level (mainly University Lille 1, Medical University (Lille 2) and the Pasteur Institute of Lille) for the period 2006-09.

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

6.2. National initiatives and cooperations

6.2.1. National initiatives

We participate in the following national projects:

- ARENA working group funded by *ACI ImpBIO*[12] (2004-2007). This national group gathers scientists (mainly biologists and computer scientists) having a common interest in RNA computational analysis.
- ANR BRASERO *Biologically Relevant Algorithms and Softwares 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 Rhones Alpes).
- *ACI ImpBIO* project REPEVOL[13] (2004-2007). The project is joint with LIRMM, *Centre d'Ecologie Fonctionnelle et Evolutive* and *Institut de Génétique Humaine* of Montpellier, and Boston University, USA. The subject of the project is the analysis of repeated structures in genomic sequences.
- *Action de Recherche Coopérative (ARC) "Optimisation de graines et indexation des banques d'ADN sur mémoire FLASH reconfigurable"* funded by INRIA (2006-2007). The project is headed by D. Lavenier (SYMBIOSE team, RU Rennes) and includes researchers from INSERM U694 (CHU Angers) and the team IP Design (LESTER, Lorient). The goal of this project is to use reconfigurable parallel computer architectures (ReMIX prototype) in order to design efficient methods of indexing and searching biological sequence data using the *multiple spaced seeds* strategy (see Sections 3.1 and 3.5.1).
- working groups *Sequence analysis* and *Structural bioinformatics* of the multidisciplinary *GDR Molecular bioinformatics*[14].
- working group *Combinatoire des mots, algorithmique du texte et du génome* of the newly created *GDR Informatique Mathématique*[15].

6.2.2. 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)
- Rennes, IRISA, Symbiose, with P. Veber and D. Lavenier, epsilon-transitions in weighted finite automata (M. Giraud)
- Evry, Laboratoire Statistique et Génome, with C. Devauchelle, A. Grossman, A. Hénaut and I. Laprevotte, alignment-free sequence comparison (M. Pupin)
- Institut de Mathématiques de Luminy, with G. Didier, local decoding of sequences (M. Pupin)

6.3. International initiatives and cooperations

6.3.1. Foreign visitors

- Daniel Brown, a professor from the University of Waterloo, Canada and currently on sabbatical in the University of California at Davis, USA, visited our team on April 2-4 and made a talk at the group seminar.

[12] http://www.lri.fr/~denise/AReNa
[13] http://www.lirmm.fr/~rivals/RESEARCH/REPEVOL
[14] http://www.gdr-bim.u-psud.fr
[15] http://www.liafa.jussieu.fr/~alp/IM.html

- A collaboration started this year with Professor Liviu Ciortuz from the Computer Science Department of the University of Iasi, Romania. He has been an invited professor of our group in summer 2006. This collaboration will give rise to a co-supervision of a master student in 2007.
- Anna Gambin and Slawomir Lasota, both associate professors at Warsaw University, stayed for three weeks with our group in August. Ewa Makosa, a master student from the same University made a 1.5-month internship in our group in July and August. (see also next Section)
- Mikhail Roytberg, senior researcher of the Institute of Mathematical Problems in Biology in Puschino (Russia), visited our team for 1.5 month in November-December and gave a talk at the group seminar. (see also next Section)

6.3.2. ECO-NET and Polonium

We currently run an ECO-NET project and a Polonium project, both funded by the French Ministry of Foreign Affairs during 2005-2006. ECO-NET is a tri-partite project, joint with russian and polish researchers, and Polonium is a bilateral french-polish cooperation.

On the russian side, the main partner is the Institute of Mathematical Problems in Biology in Puschino, and more specifically the group of M. Roytberg with which we have an active collaboration for the last two years. The main subject of the collaboration is the seed-based similarity search, both in DNA (Section 3.1) and proteins (Section 3.1.3).

On the polish side, we collaborate with the bioinformatics group at the Computer Science department of Warsaw University (J. Tiuryn, A. Gambin). Two topics have been developed within this collaboration: one on the analysis of transposable elements in plan genomes, and in particular in the *Medicaco Trancatula* genome, and another on protein seeds (Section 3.1.3), with application to mass spectrometry.

6.3.3. Bilateral cooperations

- Belgium, *Université Libre de Bruxelles, Service de conformation des macromolécules biologiques et de bioinformatique*, headed by S. Wodak and J. van Helden: inference of over-represented patterns in the regulatory regions of eukaryotic organisms. Regular meetings and student exchanges. (H. Touzet, M. Defrance)
- Canada, Université de Montréal, with M. Csűrös: seed-based indexing of genomic sequences (G. Kucherov, L. Noé), with N. El Mabrouk and J.-E. Duchesnes: RNA analysis (M. Giraud)
- Israël, Haifa University, Computer Science Department, with G. Landau, D. Hermelin: string matching and RNA modelling (G. Kucherov, H. Touzet)
- Russia, Moscow University, with R. Kolpakov: combinatorics of repetitions in words, tandem repeats in DNA sequences and `mreps` software (G. Kucherov)
- Boston University, with Prof. Gary Benson: REPEVOL project of the ACI IMPBio, integration of `mreps` to the TRDB system; Brooklyn College, CUNY, with Prof. Dina Sokol: joint work (G. Kucherov)
- London, King's College, with K. Iliopoulos: string prosessing (G. Kucherov)

7. Dissemination

7.1. Organization of workshops and seminars

7.1.1. Moscow workshop on algorithms in bioinformatics

Within our ECO-NET cooperation, we organized on July 11-13, 2006 a Workshop *Algorithms in bioinformatics*[16] that was held in Moscow at the French-Russian J.-V. Poncelet laboratory. Note that this laboratory, initially affiliated with CNRS, is now joined by INRIA.

[16] http://www.mccme.ru/albio

7.1.2. Gent-Lille workshop

Jointly with IRI (Interdisciplinary Research Institute, Lille) and VIB (Flanders Interuniversity Institute for Biotechnology, Gent), we organized a cross-border workshop devoted to bioinformatics and computational biology[17] (20/06/2006, 50 participants).

7.1.3. GTGC workgroup

J.-S. Varré is one of the committee members of the national GTGC working group[18] (Comparative Genomics Working Group) created in 2005. The group organizes two seminar sessions per year on comparative genomics. A large number of presentations are devoted to biological problems.

7.1.4. IEMN – LIFL – IRI seminar series

Since 2003 we organize joint seminars with researchers coming from IRI (Interdisciplinary Research Institute, Lille), IEMN (Electronic, Microelectronic and Nanotechnology Institute) and LIFL. The goal of those seminars is to share and exchange on problems that are at the junction of physics, mathematics, computer science and bio-informatics. The program of future and past seminars may be found at http://www.lifl.fr/BIOINFO/seminaires0506.html.

7.1.5. Journées au vert

On June 22-23, 2006, we organized a team two-days seminar in Bollezeele (Nord) in order to discuss current and future research projects carried out in the group.

7.2. Editorial and reviewing activities

- Editorial Board of BMC Algorithms for Molecular Biology (G. Kucherov)
- Program committee of ECCB 2006 (G. Kucherov), JOBIM 2006 (H. Touzet), PSI 2006 (G. Kucherov), CPM 2007 (G. Kucherov), JOBIM 2007 (G. Kucherov, H. Touzet).
- Reviewer for the journals Bioinformatics (G. Kucherov) BMC Algorithms for Molecular Biology (H. Touzet), BMC Bioinformatics (M. Giraud, H. Touzet, J.-S. Varré), Information and Computation (G. Kucherov) Information Processing Letters (H. Touzet), Nucleic Acids Research (H. Touzet), Theoretical Computer Science (G. Kucherov)
- Reviewer for the conferences CPM 2006 (H. Touzet), JOBIM 2006 (J.-S. Varré, M. Pupin), MFCS 2006 (G. Kucherov), RECOMB 2007 (J.-S. Varré), STACS 2006 (G. Kucherov, J.-S. Varré), WABI 2006 (M. Giraud).

7.3. Miscellaneous activities

- Jury of the HDR these of M. Raffinot (G. Kucherov, *rapporteur*), PhD theses of M. Rao (G. Kucherov), S. Djebali and P. Peterlongo (G. Kucherov, *rapporteur*)
- Scientific committee of the french ministry program ANR (H. Touzet)
- G. Kucherov, jointly with D. Sokol (Brooklyn College, CUNY), has been assigned to write an entry on the algorithms for approximate tandem repeats for the Encyclopedia of Algorithms, to be published by Springer Verlag in 2007.
- M. Giraud, jointly with the Symbiose project (INRIA Rennes), coordinated an exposition for the yearly event *Fête de la Science* in october 2006. Three bioinformatics puzzles (sequence assembly, motif discovery and protein classification) were presented in the *Jardin du Luxembourg* of Paris.

7.4. Meetings attended and talks

7.4.1. International Conferences

[17] http://www.iri.cnrs.fr/bn/workshop/workshop_20june.html
[18] http://biomserv.univ-lyon1.fr/~tannier/GTGC/

- CPM 2006, Combinatorial Pattern Matching, Barcelona, Spain, July 2006 (A. Liefooghe, H. Touzet, J-S. Varré [21])
- CIAA 2006, Conference on Implementation and Application of Automata, Taipei, Taiwan, August 2006 (M. Giraud [19])
- SPIRE 2006, String Processing and Information Retrieval, Glasgow, Scotland, October 2006, (H. Touzet [16])
- French-Indian Computer Science Workshop, Bangalore, India, February 2006 (G. Kucherov)
- Haifa Annual International Stringology Workshop, Israel, May 2006 (G. Kucherov)

7.4.2. National Conferences

- JOBIM 2006, *Journées Ouvertes Biologie Mathématique Informatique Biologie*, Bordeaux, July 2006 (S. Caboche, L. Noé, M. Pupin, M. Defrance)

7.4.3. Talks, meetings, seminars

- *Analyse comparative pour l'étude des gènes d'ARN*, ARENA workshop, Toulouse, December 2005 (H. Touzet)
- *Classification d'ARN: codant / non-codant*, ARENA workshop, Toulouse, December 2005 (A. Fontaine)
- *Finding regulatory elements shared by a set of genes*, IRISA, Rennes, January 2006 (M. Defrance)
- *Localisation à grande échelle de motifs nucléiques décrits par des matrices position-poids*, IRISA, Rennes, January 2006 (A. Liefooghe)
- *Combinatorial search on graphs motivated by bioinformatics applications: a case study and generalizations*, seminar at Moscow Independent University, March 2006 (G. Kucherov)
- *Studying tumor architectures using genome rearrangement theory on end-sequence profiling data*, ACI VicAnne/ARC MOCA worshop, Lille, March 2006 (J.-S. Varré)
- *Recherche de similarités dans les séquences génomiques: modèles et algorithmes pour la conception de graines efficaces*, MAB seminar, Montpellier, March 2006, and MIG seminar, Jouy en Josas, March 2006 (L. Noé)
- *Recherche de motifs par automates sur FPGA*, LINA seminar, Nantes, April 2006, and LERIA seminar, Angers, May 2006 (M. Giraud)
- *Analysis of regulatory sequences*, Gent-Lille workshop on computational biology, June 2006 (M. Defrance, A. Liefooghe, H. Touzet, JS. Varré)
- *RNA comparative analysis : structure prediction and gene prediction*, Gent-Lille workshop on computational biology, June 2006 (A. Fontaine, H. Touzet)
- *Spaced seeds for homology search*, Gent-Lille workshop on computational biology, June 2006 (G. Kucherov)
- *Modèles combinatoires pour l'analyse de structures d'ARN*, Forum des Jeunes Mathématiciennes – Mathématiques et Interactions, October 2006 (H. Touzet)
- *Application bio-informatique: gènes à protéines et gènes à ARN*, Grid'5000 seminar, Lille, October 2006 (A. Fontaine, H. Touzet)
- *Application des méthodes de réarrangements génomiques à la comparaison génomes de tumeur/génomes sains*, GTGC workshop, Nantes, October 2006 (J.-S. Varré)
- *Combinatorial search for bioinformatics*, Seminar of Computer Science Department of King's College, London, October 2006 (G. Kucherov)
- *Décodage local et application à l'alignement multiple de séquences d'ADN*, IRISA, Rennes, December 2006 (M. Pupin)

7.5. Teaching activities

Our research work finds also its expression in a strong commitment in pedagogical activities at the University Lille 1. For five 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's degree) in computer science and biology (master protéomique, master biologie-santé, master génie cellulaire et moléculaire, master interface physique-chimie, master bioinformatique) in an Engineering School (Polytech'Lille), as well as in permanent education (for researchers, engineers and technicians).

7.5.1. Invited lectures on bioinformatics

- *Non-coding RNAs*, technical session of the INSERM workshop 166 (H. Touzet)
- *Ethics and bioinformatics*, DU ethics and biomedical research, UCL, one-day session (H. Touzet)

7.5.2. Lectures on bioinformatics, University of Lille 1

- Organization of a lecture series on *Algorithms and computational biology*, master in computer science (M2), 17h (M. Pupin, H. Touzet, G. Kucherov, M. Giraud)
- *Regulatory regions analysis, Transcriptome*, master in biology (M2), one-day session (H. Touzet)
- *Computational biology*, master in computer science (M1), 50h (H. Touzet, together with C. Abbadie)
- *Bioinformatics*, master génomique et protéomique (M1), 64h (M. Pupin, J.-S. Varré)
- *Bioinformatics*, master génomique et microbiologie (M1), 40h (L. Noé)
- *Bioinformatics*, master protéomique (M2), 30h (M. Defrance, M. Pupin)
- *Bioinformatics*, master génie cellulaire et moléculaire (M2), 40h (M. Pupin, J.-S. Varré)
- *Bioinformatics*, master biologie-santé (M2), 14h (M. Pupin)
- *Bioinformatics*, master from Polytech'Lille, 24h (M. Pupin with S. Janot)
- *Bioanalysis*, master bioinformatique (M2), 34h (M. Pupin)

7.5.3. Teaching in computer science, University of Lille 1

- *Algorithmics*, first year IUT students, 40h (A. Fontaine)
- *Computers architecture*, first year IUT students, 24h (A. Fontaine)
- *Algorithmics and programming*, first year of bachelor, 120h (M. Dauchet)
- *Web technologies*, second year of bachelor, 36h (M. Defrance)
- *Automata and Languages*, second year of bachelor, 28h (A. Liefooghe)
- *Programming (Ocaml, Prolog)*, third year of bachelor, 48h (L. Noé)
- *Networks*, third year of bachelor, 72h (L. Noé)
- *Algorithmics*, third year of bachelor, 57.5h (J.-S. Varré)
- *Software project*, third year of bachelor, 35h (J.-S. Varré)
- *Object oriented programming*, third year of bachelor, 45,5h (J.-S. Varré)
- *Professional project*, first year of master, 20h (L. Noé, M. Pupin)
- *Operating systems architecture*, first year of master, 42h (L. Noé)
- *Business intelligence*, first year of master, 40h (A. Liefooghe)
- *Web technologies*, doctorate, 21h (J.-S. Varré)

7.6. Administrative activities

- Head of the graduate school in engineering sciences of the University of Lille 1 (M. Dauchet)
- Board of the SFBI, French Society of Bioinformatics (H. Touzet)
- 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)
- Member of the LIFL Laboratory council (H. Touzet)
- Head of PPF bioinformatics, created in 2005 (M. Dauchet)
- Member of the *Commission des Spécialistes* of the University Lille 1 since 2003 (J-S. Varré)
- Supervisor of the Master of Bioinformatics of the University Lille 1 (M. Pupin)

8. Bibliography

Major publications by the team in recent years

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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_uid

[8] 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.

Year Publications

Doctoral dissertations and Habilitation theses

[9] M. DEFRANCE. *Algorithmes pour l'analyse de régions régulatrices dans le génome d'eucaryotes supérieurs*, Ph. D. Thesis, Université des Sciences et Technologies de Lille, December 2006, http://www.lifl.fr/~defrance/these.pdf.

Articles in refereed journals and book chapters

[10] 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.

[11] 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.

[12] 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.

[13] S. TEMPEL, M. GIRAUD, D. LAVENIER, I.-C. LERMAN, A.-S. VALIN, I. COUÉE, A. E. AMRANI, J. NICOLAS. *Domain organization within repeated DNA sequences: application to the study of a family of transposable elements.*, in "Bioinformatics", vol. 22, n° 16, 2006, p. 1948–1954, http://dx.doi.org/10.1093/bioinformatics/btl337.

[14] H. TOUZET. *Comparative analysis of RNA genes: the CaRNAc software*, N. BERGMAN (editor). , in press, vol. Methods in Molecular Biology, Special issue on comparative genomics I, Humana Press, 2006.

[15] H. TOUZET. *Comparing similar ordered trees in linear-time*, in "Journal of Discrete Algorithms", 2006, http://dx.doi.org/10.1016/j.jda.2006.07.002.

Publications in Conferences and Workshops

[16] 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/.

[17] S. CABOCHE, V. LECLÈRE, P. JACQUES, M. PUPIN, G. KUCHEROV. *Database and comparison of non ribosomal peptides*, in "7th Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)", 2006, http://cbi.labri.fr/jobim2006/presentations/050706/am/7_Caboche.pdf.

[18] J. COHEN, F. FOMIN, P. HEGGERNES, D. KRATSCH, G. KUCHEROV. *Optimal Linear Arrangement of Interval Graphs*, in "Proceedings of the 13th International Symposium on Mathematical Foundations of Computer Science (MFCS 2006), High Tatras (Slovakia), August 28 - September 1, 2006", Lecture Notes in Computer Science, vol. 4162, Springer Verlag, 2006, p. 267–279, http://www.springerlink.com/content/wrk8346657167528/?p=a1f283d79ac14affb4d165501b84e569&pi=23.

[19] M. GIRAUD, P. VEBER, D. LAVENIER. *Path-Equivalent Removal of Epsilon-Transitions*, in "11th International Conference on Implementation and Application of Automata (CIAA)", Lecture Notes in Computer Science, vol. 4094, Springer Verlag, August 2006, p. 23–33, http://dx.doi.org/10.1007/11812128_4.

[20] V. LECLÈRE, S. CABOCHE, M. PUPIN, G. KUCHEROV, P. JACQUES. *NORINE: a recent database highlighting a large biodiversity among NRPS peptide structures and activities*, in "RSC conference, Chemical Biology: directing biosynthesis (poster)", 2006, http://www.lifl.fr/~caboche/Publications/Cambridge.pdf.

[21] 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/.

References in notes

[22] 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.

[23] 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.

[24] 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.

[25] M. CSÜRÖS, B. MA. *Rapid homology search with neighbor seeds*, in "Algorithmica", to appear, 2006.

[26] F. E, B. JP, P. GARDNER. *Exploring genomic dark matter: a critical assessment of the performance of homology search methods on non-coding RNA*, in "To appear in Genome Research", 2006.

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

[28] 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.

[29] M. FARACH-COLTON, Y. HUANG, J. WOOLFORD. *Discovering temporal relations in molecular pathways using protein-protein interactions*, in "Proceedings of the 8th Annual International Conference on Computational Molecular Biology (RECOMB'04), San Diego, California, USA, March 27-31, 2004", ACM Press, 2004, p. 150–156.

[30] M. FARACH-COLTON, G. M. LANDAU, S. CENK 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.

[31] P. GARDNER, R. GIEGERICH. *A comprehensive comparison of comparative RNA structure prediction approaches*, in "BMC Bioinformatics", vol. 5(140), 2004, http://www.binf.ku.dk/~pgardner/bralibase/bralibase1.html.

[32] 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/browse.shtml.

[33] 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.

[34] D. KAROLCHIK, R. BAERTSCH, M. DIEKHANS, T. FUREY, A. HINRICHS, Y. LU, K. ROSKIN, M. SCHWARTZ, C. SUGNET, D. THOMAS, R. WEBER, D. HAUSSLER, W. KENT. *The UCSC Genome Browser Database.*, in "Nucleic Acids Res", vol. 31, n^{o} 1, 2003, p. 51-4.

[35] 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, July 1 2003, p. 3672-3678.

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

[37] M. LI, M. MA, L. ZHANG. *Superiority and Complexity of the Spaced Seeds*, submited to Algorithmica, 2006.

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

[39] 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.

[40] 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.

[41] L. NOÉ, G. KUCHEROV. *Improved hit criteria for DNA local alignment*, in "BMC Bioinformatics", vol. 5, n^{o} 149, 14 October 2004.

[42] 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.

[43] K. PRUITT, T. TATUSOVA, D. MAGLOTT. *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins*, in "Nucleic Acids Res", vol. 33, n^{o} Database issue, 2005, p. D501-4.

[44] Y. SUN, J. BUHLER. *Choosing the best heuristic for seeded alignment of DNA sequences*, in "BMC Bioinformatics", vol. 7, n^{o} 133, march 2006, http://www.biomedcentral.com/1471-2105/7/133/.

[45] 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.

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

[47] 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.

[48] 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.

[49] S. WHELAN, P. I. W. DE BAKKER, E. QUEVILLON, N. RODRIGUEZ, N. GOLDMAN. *PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees*, in "Nucleic Acids Research", vol. 34, 2006, p. Database issue D327-D331.