INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team sequoia

Algorithms for large-scale sequence analysis for molecular biology

Futurs

THEME BIO

Activity Report

2007

Table of contents

1. Team

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

Head of project-team
Gregory Kucherov [DR CNRS, HdR]

Vice-head of project-team
Hélène Touzet [CR CNRS, HdR]

Administrative assistant
Axelle Magnier [INRIA, until October 2007]
Sandrine Catillon [INRIA, from October 2007]

INRIA project-team staff
Jesper Jansson [CR INRIA, from October 2007]

CNRS project-team staff
Mathieu Giraud [CR CNRS]

University project-team staff
Laurent Noé [MC, Université Lille 1]
Maude Pupin [MC, Université Lille 1]
Jean-Stéphane Varré [MC, Université Lille 1]

Post-doctoral fellows
Mathieu Defrance [until August 2007, now post-doc at the *Université Libre de Bruxelles*]
Sylvain Guillemot [ATER, Université Lille 1, from September 2007]
Alban Mancheron [PostDoc, INRIA, from September 2007]

PhD students
Ségolène Caboche [INRIA/Region fellowship, from October 2006]
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]
Azadeh Saffarian [MESR fellowship, from November 2007]

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

Invited researchers
Liviu Ciortuz [University of Iasi, Romania, July 1 – July 31]
Roman Kolpakov [Moscow University, Russia, September 1 – September 30]

Internships
Thomas Faisca [Université Lille 1, 01/02/2007 – 31/05/2007]
Ankit Goyal [IIT Rorkee, 01/06/2007 – 30/07/2007]
Arnaud Jacquemin [Université Lille 1, 01/09/2007 – 05/10/2007]
Sébastien Le Maguer [Université Lille 1, 15/06/2007 – 31/07/2007]
Adeline Lepaul [Université Lille 1, 01/02/2007 – 30/06/2007]
Cédric Molendi-Costes [Université Lille 1, 01/02/2007 – 30/06/2007]
Przemyslaw Piechowak [Poznan University, 01/06/2007 – 30/09/2007]

2. Overall Objectives

2.1. Introduction

Keywords: *RNA structures*, *algorithmics*, *bioinformatics*, *comparative genomics*, *computational biology*, *discrete algorithms*, *genomic sequences*, *high-performance computing*, *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.

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 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

- Year 2007 has been the year of official creation of SEQUOIA. Several members joined the project-team, some of them being recruited by INRIA.
- Our work on NORINE, published this year in *Nucleic Acids Research* journal [10], has been awarded a best poster prize at the congress of *Société Française de Microbiologie* [18]. It also constituted an important premise for setting up an European project proposal.
- Work [11] that appeared in the high-standard biological journal *Trends in genetics*, brings a novel insight into the fundamental issue of *statistical significance* of word appearance in a genome.
- We have been invited to present our work on RNA structure inference in a chapter of a popularizing book devoted to comparative genomics [15]. This book provides a collection of robust protocols for molecular biologists studying comparative genomics, and is intended to become a major resource in this field.

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 [25], 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 [48]. 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 [30], [49], [67], [44], [34], [66], [41] 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 [54] [9] 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 3.6.3) 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], [33], [53].

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 [43] (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. [56]) 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 ([29] 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 [46], 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: *RNA*, *base pairings*, *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, 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.

3.3.1. 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 [55], [60], [15] (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], [61], [62]. 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 [35]. 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 [59].

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 [32], [40]. 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 [31]. 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 [64] 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, 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 [45], [50].

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 [26] 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 6.1).

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 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 3.3).

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 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 3.3).

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 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*, *subset seeds*, *transition constrained seeds*.

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

YASS [54] [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. Besides the Web-server of our team, version 1.12 of YASS is available from the INRIA software web page [2].

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

[2] http://www.inria.fr/valorisation/logiciels/vie.fr.html

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.

We recently developed a more general tool for designing subset seeds according to its general definition. This tool named IEDERA implements our recent work of [23]. The first release is now available at http://bioinfo.lifl.fr/yass/iedera.php.

4.3. Noncoding RNA tools

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

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 [37], and is presented in a book chapter devoted to comparative genomics [15]. GARDENIA is a new 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 [28]. GARDENIA appears to be more robust than similar existing programs, such as those of the Vienna Package.

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

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]. The server uses pre-computed background models for Human, Mouse and Rat genomes derived from annotated genes with REFSEQ identifiers [57] available from the UCSC Genome Browser assembly [42] (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 has been released in August 2006, and has been used by several biology research groups [65], [52], [63]. The TFM-Scan program implements an efficient algorithm for the location of PWM matrices on a sequence [8]. TFM-Pvalue is a new program to compute score thresholds for PWMs. This computation is a prerequisite to various applications of PWMs, which opens a broad range of applications of the TFM-Pvalue algorithm. Note that our algorithm outperforms existing solutions both in terms of accuracy and efficiency. According to our experimental evaluation against the algorithm proposed in [27], we are able to obtain a result with the same accuracy more quickly or to obtain a more accurate result within the same computation time.

4.5. Protea

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

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 5.4.2. 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.

4.6. Norine

Keywords: *database*, *nonribosomal peptide synthesis*.
URL: http://bioinfo.lifl.fr/norine

We continue to develop a database of NRPS peptides called NORINE [3]. This is a unique resource as there has been no centralized depository of these data before. Among existing related resources, NRPS-PKS [4] is focused on the synthases and contains only a very limited number of peptides, other resources like PubChem [5] or ChEBI [6] 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. Today, NORINE contains more than 700 peptides, described in about 350 publications. 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 data has been completed: a great deal of new entries and new annotations have been added. On the other hand, NORINE has been extended by several new tools: the user can now search the database for a given molecule or for a given structural pattern, as well as for molecules with a given monomeric composition. We also added new tools for a 2-D visualization of peptide structure and an editor for drawing a peptide structure in order to obtain its graph representation.

4.7. 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 3.2), 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 distributed from the INRIA free software server [10].
- `grappe` (http://www.lifl.fr/~kucherov/software/grappe) 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 3.4).
- 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 clustering with associated probability for each node of the clustering. The core of the HUGO algorithm is based on graph-theoretic techniques.

[3] **no**n-**ri**bosomal peptides, with **ine** as a typical ending of names of nonribosomal 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

5. New Results

5.1. Sequence similarity and repetitions

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

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 [46], 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) [23]. 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 4.2). An extended journal version of this paper has been submitted 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 [46] 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 [51], [47]), 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.

A paper describing these studies is still under preparation. The PhD work of Marta Girdea started in September with the goal to further develop these studies for new applications.

5.1.3. Computer architecture for seed-based sequence comparison

Within our ARC Flash collaboration with the SYMBIOSE team in INRIA-Rennes (see section 6.2), we designed a technology that implements subset-seed-based search for protein sequences (see previous section) in a specialized parallel hardware. The hardware embeds a reconfigurable architecture (FPGA) that is tightly connected to large capacity Flash memory chips. The efficiency of this prototype relies on a large amount of available memory. The memory stores a large redundant index with a low-latency access. Several ideas have been explored in order to reduce the size of the index.

Compared to traditional approaches represented by the BLASTP software, we obtain both a significant speed-up and better results. This has been demonstrated in a large-scale experiment on the hard-masked human chromosome 1 (UCSC Release hg18) translated with respect to the six reading frames (85×10^6 amino acids) compared against a set of seven archeae and bacteria proteomes (5.5×10^6 amino acids) deriving from a study on mitochondrial diseases.

To the best of our knowledge, this work is the first attempt to take advantage of efficient seed-based algorithms in parallel implementation of similarity search. The hardware architecture, algorithmic solutions and experimental benchmarks have been presented to the *Parallel Biocomputing Conference (PBC'07)* [24]. A journal paper describing extensions of this work is under preparation.

5.1.4. Statistics of genomic word frequencies

We continued the study, started last year, of the probabilistic distribution of words (k-mers) in genomic sequences. We proposed that the distribution of DNA words in genomic sequences is primarily characterized by a so-called double Pareto-lognormal distribution, which explains lognormal and power-law features found across all known genomes. This study shows that this distribution cannot be explained by Bernoulli or Markov models usually used to model genomic sequences, and confirms the fundamental role of duplications in the genome evolution that are not captured by those models. This work, joint work with Prof. Miklós Csűrös from the University of Montréal, appeared in the high-level journal *Trends in genetics*.

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 [16].

We also continued our work on the RNA alignment hierarchy, originally initiated in [28]. 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 extended the NP-completeness results obtained in [28] and fully solved the alignment hierarchy complexity analysis. We also carried out experimental analyses of the average complexity of some polynomial instances of the hierarchy. A paper describing these results is under revision at the international journal *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Besides the theoretical aspects of the alignment hierarchy, we also investigated its practical use. We focused on a new polynomial algorithm and introduced a well-suited combinatorial data structure, named the *arc-annotated alignment graph*. It allowed us to propose several improvements within this framework. First, we take into account dependencies between neighboring positions in the sequence and in the structure. This complex evolutionary model help to create alignments that better conform to the biology. Secondly, we add several enrichments, such as local search, constraints coming from the primary structure, speed search with k errors and multiple alignment. We are preparing a submission to the *Journal of Bioinformatics and Computational Biology* on that subject.

5.3. Cis-regulatory sequence analysis

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

5.3.1. Score threshold computation for PWMs

The usage of PWMs requires a knowledge of statistical significance of a word according to its score. This is done by defining the P-value of a score, which is the probability that the background model can achieve a score larger than or equal to the observed value. This gives rise to the following problem: Given a P-value, find the corresponding score threshold. Existing methods rely on dynamic programming or probability generating functions. For many examples of PWMs, they fail to give accurate results in a reasonable amount of time. We proved that the problem is NP-hard. Then, we described a novel algorithm that solves the P-value problem efficiently. The main idea is to use a series of discretized score distributions that improves the final result step by step until some convergence criteria is met. Moreover, the algorithm is capable of calculating the exact P-value without any error, even for matrices with non-integer coefficient values. A paper describing the algorithm is under revision for *Algorithms for Molecular Biology* [17].

5.4. Comparative genomics applications

5.4.1. Alignment-free sequence comparison

In collaboration with several people from the *Laboratoire Statistique et Génome* from Evry (see Section 6.2.2) and with Gilles Didier from the *Institut de Mathématiques de Luminy*, we worked on local decoding of sequences and alignment free comparison and its application to HIV/SIV subtyping. This work was published in two papers, one on the algorithm [2] and another on its application [12].

Accuracy of sequence alignment is necessary for constructing reliable phylogenetic trees. However, it may lead to a loss of sequence information due to the deletion of ambiguous alignment regions. We proposed an alternative method, without time-consuming alignment requirement, to perform a better sequence comparison and tree construction. In this method, each sequence position is decoded in a specific identifier defined by the N-tuple context in the sequence itself and all other sequences studied. This decoding explores an extreme case of hidden Markov modeling as it defines the sequence of states with the greatest variety such that the knowledge of N-tuples including a given position is sufficient to decode the state associated to this position. The algorithm is efficient: its complexity is linear in the total length of the set (whatever the order N is) both in time and memory space. So it could deal with huge data sets (in the length of the sequences and in their number) such as complete viral genomes.

A dissimilarity matrix calculated from the decoded sequences makes it possible to construct trees. Results obtained on HIV/SIV genome sequences, regardless of the genome complexities, were in good agreement with our best HIV/SIV taxonomic knowledge (other datasets reinforces a good accuracy). Finally, the decoded sequence blocks may also be used as anchor points in those similarity-block-based alignment programs in order to improve the alignment quality.

5.4.2. Computational identification of protein-coding sequences

Gene prediction is an essential step in understanding the genome of a species once it has been sequenced. For that, a promising direction in current research on gene finding is a comparative genomics approach. We designed a novel approach to identify evolutionary conserved protein-coding sequences in genomes. The rationale behind the method is that protein coding sequences should feature mutations that are consistent with the genetic code and that tend to preserve the function of the translated amino acid sequence. The algorithm takes advantage of the specific substitution pattern of coding sequences together with the consistency of reading frames. It has been implemented in a software called PROTEA. We have conducted a large scale analysis on thousands of conserved elements across eighteen eukaryotic genomes, including the Human genome. This experiment reveals the existence of new putative protein-coding sequences. Most of them are likely to be involved in alternative splicing transcripts, or to correspond to unannotated exons of predicted genes. This work appeared in [21].

5.4.3. 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 [38] and estimated their theoretical sensitivities. We discovered some bias in the benchmarks of [36]. Finally, following the methodology of [58], we compared the predictions on non-coding RNA candidates versus known RNA on *E.coli*. This work has been presented in [20]. We are preparing a paper on those themes.

5.4.4. Computing clusters of homologous sequences

With the availability of numerous genomic sequences, it is now a common need to compute *multi-copy* repeats occurring in one or several input sequences. In particular, this task occurs in RNA gene prediction project described in the previous section. Note that usual multiple alignment techniques usually cannot be applied because of their prohibitive computation cost. Therefore, one needs first to perform all pairwise comparisons of input sequences in order to compute all pairs of similarity regions and then to recover clusters of similar sequences out of this information.

This task was the subject of master diploma work of C. Molendi-Costes carried out this year. The goal was to develop a robust software for this task, using methods implemented in two experimental programs designed earlier in the team. This work resulted in a new software, named YOPCLUSTER, which is currently being tested.

5.5. Nonribosomal peptide synthesis

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

As presented in Section 4.6, 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. Finally, they have several important activities, such as antibiotic, anti-inflammatory, antithrombotic, antitumor, calmodulin antagonist, immunomodulating, protease inhibitor, siderophore, surfactant, and toxin. In our poster presented this year to the congress of the French Society for Microbiology [18], we showed how NORINE data can be used to explore the structural and compositional diversity of NRPS peptides.

This year, we developed and implemented in NORINE several new efficient algorithms to compare NRPS molecules represented as non-oriented labeled graphs. The first one looks up for a given molecule in the database. This amounts to testing the isomorphism of two labeled graphs which is a difficult computational problem in general. To cope with it, we applied the method of so-called *association graph* used in [39] which allowed us to obtain an efficient algorithmic solution.

Looking for an exact occurrence of a molecule in the database was then extended to the search for all molecules containing a given *structural pattern*. The latter is formalized through a subgraph with nodes labeled either by monomers or by a special joker label that can be substituted by any specific monomer. Using the association graph, we were able to obtain an efficient algorithm for this problem too.

Finally, we also designed and implemented an algorithm to look for molecules with a given monomeric composition (precisely or with a given number of "errors"). All these new features of NORINE have been presented this year in an article published in *Nucleic Acids Research* [10].

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).

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

- 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. Guillochon, 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 Prof. Philippe Jacques from ProBioGem.
- 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, 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.
- This year we started a new collaboration with F. Sebbane (INSERM U 801) on the analysis of *Yersinia pestis* genome for the discovery of small non-coding RNAs.

6.2. National initiatives and cooperations

6.2.1. National initiatives

We participate in the following national projects:

- *ACI ImpBIO* working group ARENA [12] (2004-2007). This national group gathers scientists (mainly biologists and computer scientists) having a common interest in RNA computational analysis.
- *ACI ImpBIO* project REPEVOL [13] (2004-2007). The final reports of ARENA and REPEVOL projects were presented at the concluding Workshop of the ImpBIO program on October 4-5 in the *Institut Henri Poincaré* in Paris.
- 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).
- *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, Rennes) and includes researchers from INSERM U694 (CHU Angers) and the team IP Design (LESTER, Lorient). The goal of this project was to use reconfigurable parallel computer architectures (ReMIX prototype) in order to design efficient methods of indexing and searching biological sequence data using the *multiple subset seeds* strategy (see Sections 3.2 and 3.6.1).
- 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.

[12] http://www.lri.fr/~denise/AReNa
[13] http://www.lirmm.fr/~rivals/RESEARCH/REPEVOL

- 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 *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 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: E. Rocha (*Atelier de bioinformatique*, Paris), B. Jacq (Institut de Biologie du Développement de Marseille), M. Régnier (INRIA-Rocquencourt), P. Peterlongo (INRIA-Rennes), C. Preda (Univ. Lille 2), E. Prieur (Univ. Rouen), A. Richard (INRIA Rhône-Alpes), M. Raffinot (LIAFA, Paris), F. Sebbane (INSERM, Lille).

6.3. International initiatives and cooperations

6.3.1. European projects

We are a part of the proposal of a large collaborative European project within the call *Food, Agriculture, Fisheries and Biotechnologies* of FP7 (call KBBE-2007-2A). The project is called NOVAPIC: *Novel Assembly Line Catalytic Machinery for Effective Production of Innovative Bioactive Compounds*. Our role in this project is to provide some bioinformatics tools and primarily to contribute with the NORINE database (see Sections 4.6,5.5). Maude Pupin is responsible for one of the Work packages of the project devoted to bioinformatics tasks. The project has been selected at the first evaluation stage, the full proposal will be submitted in February 2008.

6.3.2. Foreign visitors

- Professor Liviu Ciortuz from the Computer Science Department of the University of Iasi, Romania, visited our group for one month in summer 2007.
- Jesper Jansson (at the time, a JSPS Post-doctoral Fellow at Kyushu University, Japan) visited the SEQUOIA project-team during April 30–May 5, 2007, and gave a talk on May 4 entitled “Combinatorial Algorithms for Constructing Phylogenetic Supertrees”.
- Roman Kolpakov from Moscow University visited the project-team for one month in September as an invited professor of INRIA. During this visit, he collaborated with G. Kucherov on palindrome detection in biological sequences.
- Mikhail Roytberg (Institute of Mathematical Problems in Biology, Russia) made a 3-days visit to the team in November.

6.3.3. Bilateral cooperations

[14] http://www.gdr-bim.u-psud.fr

[15] http://www.gdr-im.fr/

- Belgium, *Université Libre de Bruxelles, Service de conformation des macromolécules biologiques et de bioinformatique*, headed by J. van Helden: inference of over-represented patterns in the regulatory regions of eukaryotic organisms. Regular meetings and student exchanges. (H. Touzet)
- 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)
- Poland, Warsaw University, A. Gambin, S. Lasota: seed-based search in protein sequences, transposon analysis (G. Kucherov, L. Noé), Agricultural University of Krakow, Department of Genetics, D. Grzebelus: transposon analysis (G. Kucherov)
- UK, Cambridge, Isaac Newton Institute for Mathematical Sciences, with C. Semple: phylogenetics (S. Guillemot, visit in December 2007)
- UK, London, King's College, with K. Iliopoulos, M. Crochemore: string processing (G. Kucherov)
- USA, Boston University, with Prof. G. 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)
- 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)

7. Dissemination

7.1. Organization of workshops and seminars

7.1.1. GTGC working group

J.-S. Varré is one of the committee members of the national GTGC working group [16] (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.

7.1.2. Arena working group

H. Touzet organized a national workshop on the subject of non coding RNA bioinformatics [17] on march 26-28, 2007. This multidisciplinary meeting gathered 55 researchers coming from computer science, biology and physics.

7.1.3. National meeting on algorithms, word combinatorics, and bioinformatics

H. Touzet and G. Kucherov took part in the organization of a joint meeting [18] of the working group *Sequence analysis* of the *GDR Molecular bioinformatics* and the working group *Combinatoire des mots, algorithmique du texte et du génome* of the *GDR Informatique Mathématique*. The meeting was held in Marne-la-Vallée on September 26-28, 2007, and was dedicated to the title of emeritus professor awarded to Maxime Crochemore. Other members of SEQUOIA attended the meeting and gave talks.

7.1.4. PPF Bioinformatique meeting

M. Pupin organized a one-day meeting in Lille on the 18th of June for the *PPF Bioinformatique of Lille*. This was the first time that scientists of the PPF present locally their work dedicated to bioinformatics.

[16] http://biomserv.univ-lyon1.fr/~tannier/GTGC/

[17] http://www2.lifl.fr/SEQUOIA/Arena/

[18] http://www.lifl.fr/SEQUOIA/Sequences/

7.1.5. 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 bioinformatics. The program of future and past seminars may be found at http://www.lifl.fr/SEQUOIA/seminairesIEMNLIFLIRI.html.

7.2. Editorial and reviewing activities

- Editorial Board of BMC Algorithms for Molecular Biology (G. Kucherov)
- Program committee of CPM 2007 (G. Kucherov), JOBIM 2007 (G. Kucherov, H. Touzet), CPM 2008 (G. Kucherov), CSR 2008 (G. Kucherov)
- Reviewer for the journals Bioinformatics (G. Kucherov), BMC Bioinformatics (G. Kucherov, H. Touzet, J.-S. Varré), Journal of Automata, Languages and Combinatorics (G. Kucherov), Journal of Computer and System Science (G. Kucherov), Journal of Discrete Algorithms (G. Kucherov), Mathematics in Computer Science (G. Kucherov), Nordic Journal of Computing (J. Jansson), Nucleic Acids Research (H. Touzet), Theoretical Computer Science (G. Kucherov), Theory of Computing Systems (J. Jansson)
- Reviewer for the conferences ALT 2007 (G. Kucherov), CIAA 2007 (G. Kucherov), CPM 2007 (G. Kucherov, H. Touzet, J.-S. Varré), JOBIM 2007 (G. Kucherov, H. Touzet, M. Giraud), Majec-STIC 2007 (M. Giraud), RECOMB 2007 (J.-S. Varré), WABI 2007 (G. Kucherov), STACS 2008 (H. Touzet).
- Reviewer for American Mathematical Society (AMS)'s Mathematical Reviews (MR) (J. Jansson)

7.3. Miscellaneous activities

- Jury of the PhD theses of Celine Kuttler (H. Touzet, president), Céline Meslin (M. Pupin, examinateur), Aida Ouangraoua (H. Touzet, rapporteur), Élise Prieur (H. Touzet, rapporteur), Fabrice Touzain (G. Kucherov, co-encadrant)
- Reviewers for the french ministry program ANR (G. Kucherov, H. Touzet)
- With the help of INRIA communication staff (M.-A. Enard), the Sequoia team designed two new bioinformatics puzzles (RNA structures, sequence alignments) (coordination: M. Giraud). The puzzles were presented during the yearly event *Fête de la Science / SciencesOparK* in October 2007, in Lille.

7.4. Meetings attended and talks

7.4.1. International Conferences

- ECCB 2006, European Conference on Computational Biology, Eilat, Israel, January 2007 (G. Kucherov, program committee member)
- WAW, Workshop on Algorithms on Words, Turku, Finland, March 2007 (G. Kucherov, invited speaker)
- CIAA 2007, Conference on Implementation and Application of Automata, Prague, Czech republic, July 2007 (L. Noé [23])
- ISMB/ECCB 2007, International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology, Vienna, Austria, July 2007 (M. Pupin, H. Touzet)
- MCCMB 2007, Moscow Conference on Computational Molecular Biology, Moscow, Russia, July 2007 (G. Kucherov)

- CPM, Combinatorial Pattern Matching, London, Canada, July 2007 (G. Kucherov, program committee member)
- PPAM/PPC 2007, Parallel Processing and Applied Mathematics, Workshop on Parallel Biocomputing, Gdansk, Poland, September 2007 (M. Giraud [24])
- BIBM 2007, Bioinformatics and Biomedecine Conference, Silicon Valley, California, November 2007 (A. Fontaine [21])
- BBC 2007, Benelux Bioinformatics Conference, Leuven, Belgium, November 2007 (A. Fontaine [22])
- FSTTCS 2007, International Conference on the Foundations of Software Technology and Theoretical Computer Science, and the post-conference workshop on Bioinformatics and Systems Biology, New Delhi, India, December 2007 (J. Jansson).

7.4.2. National Conferences

- 7th congress of the French Society for Microbiology, Nantes, June 2007 (S. Caboche)
- JOBIM 2007, *Journées Ouvertes Biologie Mathématique Informatique Biologie*, Marseille, July 2007 (S. Caboche, M. Defrance, A. Fontaine, M. Giraud, A. Liefooghe, M. Pupin, H. Touzet, J.-S. Varré)

7.4.3. Talks, meetings, seminars

- *Alignement de structures d'ARN*, Seminar of the Institut Gaspard Monge, Marne-la-Vallée, January 2007 (H. Touzet)
- *Sur la distribution des fréquences d'oligonucléotides dans un génome*, Seminar IRI, Lille, February 2007 (G. Kucherov)
- *Recherche de motifs ARN dans un genome, vite et bien*, ARENA workshop, Lille, March 2007 (H. Touzet)
- *Paysages énergétiques de l'ARN*, Seminar IRI, Lille, March 2007 (H. Touzet)
- *On genomic word frequencies*, London Stringology Days (LSD), London, March 2007 (G. Kucherov)
- *Comparaisons de séquences à base de graines espacées pour la recherche d'ARN non-codants*, ARENA workshop, Lille, March 2007 (M. Giraud)
- *Modèles combinatoires pour la comparaison d'ARN* Seminar of LIAFA, Université Paris VII, June 2007 (H. Touzet)
- *Friandises autour des architectures spécialisées pour la bioinformatique*, Seminar of the Institut Gaspard Monge, Marne-la-Vallée, June 2007 (M. Giraud)
- *Calcul de P-valeur efficace et exact pour un motif PWM* Journées algorithmique, combinatoire du texte et applications en bio-informatique, Marne-la-Vallée, September 2007 (J.-S. Varré)
- *Automate des graines sous-ensemble* Journées algorithmique, combinatoire du texte et applicatons en bio-informatique, Marne-la-Vallée, September 2007 (L. Noé)
- *Mes histoires d'amour avec les travaux de Maxime Crochemore* Journées algorithmique, combinatoire du texte et applicatons en bio-informatique, Marne-la-Vallée, September 2007 (G. Kucherov)
- *Algorithmes pour l'analyse de ARN non-codants: comparaison et prédiction* Seminar BIA-INRA, Toulouse, October 2007 (H. Touzet)
- *Matrices PWM : localisation à grande échelle et calcul de score seuil* Seminar of the Statistics for Systems Biology group, Jouy-en-Josas, December 2007 (J.-S. Varré)

7.5. 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).

7.5.1. Lectures on bioinformatics, University of Lille 1

- Organization of a lecture series on *Algorithms and computational biology*, master in computer science (M2), 17h (M. Pupin, J.-S. Varré, G. Kucherov, J. Jansson)
- *Regulatory regions analysis, Transcriptome*, master in biology (M2), one-day session (H. Touzet)
- *Computational biology*, master in computer science (M1), 50h (H. Touzet, S. Caboche, together with C. Abbadie)
- *Bioinformatics*, master génomique et protéomique (M1), 64h (J.-S. Varré, 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)
- *Bioinformatics*, master from Polytech'Lille, 24h (M. Pupin with S. Janot)

7.5.2. 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)*, second year of bachelor, 36h (M. Pupin)
- *Algorithmics*, third year of bachelor, 25h (A. Liefooghe)
- *Programming (Ocaml, Prolog)*, third year of bachelor, 48h (L. Noé)
- *Programming (C)*, third year of bachelor, 36h (L. Noé)
- *Networks*, third year of bachelor, 36h (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)
- *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)

7.6. Administrative activities

- 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 – University Lille 1, since 2007 (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*) 2007 of INRIA-Rennes - Bretagne Atlantique (G. Kucherov)
- Member of hiring commitee for the recruitment of a CNRS bioinformatics engineer in Strasbourg (H. Touzet)
- Member of hiring committee for the financial and administrative manager of INRIA Lille (M. Pupin)

8. 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.

Year Publications

Articles in refereed journals and book chapters

[10] S. CABOCHE, M. PUPIN, V. LECLÈRE, A. FONTAINE, P. JACQUES, G. KUCHEROV. *NORINE: a database of nonribosomal peptides*, in "Nucleic Acids Research", 2007, http://nar.oxfordjournals.org/cgi/content/abstract/gkm792?ijkey=QQrD7uHNr8uBJp4&keytype=ref.

[11] M. CSÜRÖS, L. NOÉ, G. KUCHEROV. *Reconsidering the significance of genomic word frequencies*, in "Trends in Genetics", vol. 23, n^{O} 11, November 2007, p. 543–546, http://dx.doi.org/10.1016/j.tig.2007.07.008.

[12] G. DIDIER, L. DEBOMY, M. PUPIN, M. ZHANG, A. GROSSMANN, C. DEVAUCHELLE, I. LAPREVOTTE. *Comparing sequences without using alignments: application to HIV/SIV subtyping*, in "BMC Bioinformatics", 2007, http://www.biomedcentral.com/1471-2105/8/1.

[13] M. GIRAUD, P. VEBER, D. LAVENIER. *Path-Equivalent Developments in Acyclic Weighted Automata*, in "International Journal of Foundations of Computer Science", vol. 18, n^{O} 4, 2007, p. 799-812, http://www.lifl.fr/~giraud/publis/gvl-ijfcs-07-preprint.pdf.

[14] D. GRZEBELUS, L. LASOTA, T. GAMBIN, G. KUCHEROV, A. GAMBIN. *Diversity and structure of PIF/Harbinger-like elements in the genome of Medicago truncatula*, in "BMC Genomics", vol. 8, n^{O} 409, 9 November 2007, http://www.biomedcentral.com/1471-2164/8/409/.

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

[16] H. TOUZET. *Comparing similar ordered trees in linear-time*, in "Journal of Discrete Algorithms", vol. 5, 2007, p. 696-705, http://dx.doi.org/10.1016/j.jda.2006.07.002.

[17] H. TOUZET, J.-S. VARRÉ. *Efficient and accurate P-value computation for Position Weight Matrices*, in "Algorithms for Molecular Biology", (in press), 2007.

Publications in Conferences and Workshops

[18] S. CABOCHE, V. LECLÈRE, M. PUPIN, G. KUCHEROV, P. JACQUES. *Norine: une nouvelle base de données qui met en exergue la biodiversité des structures et des activités des peptides synthétisés par la voie non-ribosomale (NRPS)*, in "septième congrés de la Société Francaise de Microbiologie (SFM)", (poster), 2007.

[19] J.-E. DUCHESNE, M. GIRAUD, N. E. MABROUK. *Seed-based exclusion method for non-coding RNA gene search*, in "International Computing and Combinatorics Conference (COCOON)", Lecture Notes in Computer Science, vol. 4598, Springer, 2007, p. 27-39, http://www.lifl.fr/~giraud/publis/dgem-cocoon-07-lncs.pdf.

[20] 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.

[21] A. FONTAINE, H. TOUZET. *Computational identification of protein-coding sequences by comparative analysis*, in "Proceedings of the 1st IEEE international conference on Bioinformatics and Biomedecine (BIBM), Silicon Valley, California", IEEE Computer Society, 2007, p. 95–102.

[22] A. Fontaine, H. Touzet. *Identification of protein-coding genes and RNA genes by comparative analysis*, in "Benelux Bioinformatics Conference (BBC)", (poster), 2007.

[23] 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/.

[24] P. Peterlongo, L. Noé, D. Lavenier, G. Georges, J. Jacques, G. Kucherov, M. Giraud. *Protein similarity search with subset seeds on a dedicated reconfigurable hardware*, in "Proc. of the Workshop on Parallel Computational Biology (PBC), September 9-12 2007, Gdansk (Poland)", Lecture Notes in Computer Science, to appear, Springer Verlag, 2007.

References in notes

[25] 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.

[26] 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.

[27] M. Beckstette, R. Homann, R. Giegerich, S. Kurtz. *Fast index based algorithms and software for matching position specific scoring matrices*, in "BMC Bioinformatics", n^{o} 7, 2006.

[28] 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/ .

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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.

[35] E. FREYHULT, J. BOLLBACK, 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.

[36] 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.

[37] 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.

[38] 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.html.

[39] M. HATTORI, Y. OKUNO, S. GOTO, M. KANEHISA. *Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways*, in "Journal of American Chemical Society", vol. 125, 2003, p. 11853-65.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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", Database issue, vol. 33, 2005, p. D501-4.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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/.

[64] 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.

[65] 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.

[66] 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.

[67] 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.