INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team Adage

Applying Discrete Algorithms to GEnomics
Algorithmique Discrète et ses Applications à la GÉnomique

Lorraine

THEME BIO

Table of contents

1. Team

ADAGE *is a project-team of LORIA (UMR 7503) affiliated with CNRS, INRIA,* HENRI POINCARÉ *University of Nancy 1, University of Nancy 2, and INPL.*

Head of project-team
Grégory Kucherov [CR INRIA]

Administrative assistant
Céline Simon [TR INRIA]

Research scientists
Isabelle Debled-Rennesson [Maître de conférences, IUFM de Lorraine, *détachée* CR INRIA from September 2003]
Jean-Luc Rémy [assigned to syndical activities within 30% of annual service, CR CNRS]

PhD students
Laurent Noé [grant MJENR]
Fabrice Touzain [grant INRIA co-sponsored by the Lorraine region]

Post-doctoral fellow
Alexey Vitreschak [INRIA]

Invited professor
Mikhail Roytberg [Institut for Mathematical Problems in Biology, Russia, May-July 2004]

Internships
Franck Rapaport [DEA program of Nancy, Supelec Metz, February-August 2004]
Trung Nguyen [Master program of ENS Paris, July-August 2004]
Steven Corroy [ENSIMAG student, July-August 2004]

External members
Jocelyne Rouyer [*retraitée*]

2. Overall Objectives

The project-team ADAGE was created on January 1, 2001, as a result of the evolution of the POLKA project-team. The general goal of ADAGE is to develop efficient algorithms on discrete structures (such as words, trees, polyominoes, ...). This goal leads us to study in depth mathematical properties of those structures, that can be of combinatorial or probabilistic nature.

One of our research directions is *word combinatorics and sequence algorithms*. Here, we work on the complexity analysis of problems on words (texts, or symbolic sequences) and on the development of efficient algorithms on words. Another research direction belongs to the area of *discrete geometry*. The structures studied here are discrete geometric objects, described by sets of points in $\mathbb{Z}^2$ or $\mathbb{Z}^3$. As in the previous case, our goal is to develop efficient algorithms that either verify some properties or that compute some geometric parameters of those structures.

Often, we need to study our models from a probabilistic point of view in order to estimate their "typical" properties or their accuracy on typical data. We then get interested in a probabilistic analysis of the underlying model.

One application area of our models and algorithms is of a particular importance to us: this is computational biology, where discrete models come up in a very natural and essential way. Here, we are carrying out a number of projects on DNA sequence analysis. Those problems essentially use biological knowledge and are mostly done in collaboration with biologists.

We give a special attention to implementing our algorithms into experimental software systems and to making them available to the scientific community. Two deliverable DNA sequence analysis programs are currently being developed by our team: the first one, called *mreps*, allows to compute all tandem repeats in

a given DNA sequence; another one, called YASS, computes all similarity regions between two genomic sequences or within a single one. Another sequence analysis software, named *grappe*, was developed earlier.

3. Scientific Foundations

Keywords: *algorithmic complexity*, *discrete algorithms*, *discrete geometry*, *discrete structures*, *sequence algorithms*, *string matching*.

If we define the research area of our project-team by "stepwise refinement", the first step would be to assign it to the area of *discrete algorithms*. Constructing a discrete model of a real-world phenomenon means, in mathematical terms, representing it through a *discrete structure*, such as graphs, words, trees, a set of points in a space, etc. To use discrete structures, we have to study their properties. As computer scientists, we are primarily interested in *algorithmic properties*, in particular in the *efficiency* or the *complexity* of involved computations.

In order to develop efficient algorithms on discrete structures and to analyze and optimize those algorithms, we have to understand thoroughly the properties of underlying structures. These properties can be *combinatorial* (or exact) or *probabilistic* (statistical, or typical), depending on whether the underlying model is defined deterministically or probabilistically.

To be more specific, we are carrying out fundamental studies in the following research areas.

3.1.1. Text algorithms

The area of string algorithms (also called text or sequence algorithms) has been very actively developed during last years, as witnessed by the publication of several monographs [39][43][37][38]. While string algorithms remain a natural part of discrete algorithms in general, they form now their own research area, similar to graph algorithms for example. Recent advances in string algorithms have been motivated by their numerous applications, of which the computational biology and the web search are two most salient examples. Our general goal here is to develop new efficient algorithms on words, based on our studies of word combinatorial properties. A direct application of those algorithms is the analysis of biological sequences, that we will discuss in Section 4.1.

3.1.2. Discrete geometry

While words are general discrete structures, here we are interested in discrete objects having a geometric (planar or spatial) interpretation and studied within the area of *Disrete Geometry*. Its general goal is to define a theoretical framework to translate to $\mathbb{Z}^n$ basic notions of the Euclidean geometry (such as distance, length, convexity, ...) as "faithfully" as possible. Several approaches exist to pursue this goal [35]. In our studies, we follow an arithmetical approach, where discrete objects, as straight lines or planes, are defined with arithmetical defnitions. These analytical definitions allow us to represent in a compact way any elementary digital object, to study some objects that are intrinsically discrete (and are not only approximations of continuous objects), and to define infinite discrete objects.

Methods of discrete geometry are mainly applied to geometric and graphical information, in particular to image and document processing and to medical imaging. However, other application areas exist, such as the cristallography for example. In general, this research direction is in fast progress now, as it is witnessed by the international conference *Discrete Geometry for Computer Imagery*. A technical committee on discrete geometry (TC18) of International Association of Pattern Recognition (IAPR) has been created[1] in order to promote this research area.

3.1.3. Discrete probability

Probabilistic models and probabilistic analysis are getting an increasing importance in our studies in general, and in bioinformatics applications in particular (see Section 4.1). Our contribution here is of applicative nature, as we develop, study, or use specific probabilistic models in order to solve our bioinformatics problems.

[1] http://www.cb.uu.se/~tc18/

4. Application Domains

4.1. Bioinformatics

Keywords: *DNA sequence*, *bioinformatics*, *biology*, *computational biology*, *gene*, *promoter*, *sequence alignment*.

4.1.1. Introduction

Discrete models come up virtually in all application areas but one of them plays to us a particular role: this area is molecular biology that studies biological macromolecules – DNA, RNA and proteins. In general, we are interested in the linear structure of these molecules. In other words, we are interested in "fingerprints" of biological phenomena in nucleic or protein sequences. Those fingerprints are described in terms of *patterns* (or *motifs*), and one of our main objectives is to identify, search and analyze those motifs using methods of discrete algorithmics and probabilistic analysis.

We now present research projects in bioinformatics that we are currently carrying out in our team. Most of them are done in collaboration with groups of biologists and focus on the sequence level, trying to apply our knowledge of sequence analysis methods, gained in theoretical studies. However, some of those projects, described in Section 4.1.5, go beyond a pure sequence analysis and try to study the spatial structure of DNA molecules.

4.1.2. Promoter analysis of bacterial genomes

Some sites in the non-coding part of the genome are directly involved in the transcription regulation. The knowledge of those sites would allow us to identify co-regulated genes, to determine associated regulatory mechanisms and to possibly identify proteins with unknown functions. In the framework of the theme *Bioinformatique et applications à la Génomique* of the *Pôle de Recherche Scientifique et Technologique (PRST) Intelligence Logicielle*, we work on the identification and classification of regulatory sites in the *Streptomyces coelicolor* bacterium, in collaboration with scientists of the *Laboratoire de Génétique et de Microbiologie de l'Université Henri Poincaré de Nancy* (Pierre Leblond, Bertrand Aigle). Note that this bacterium presents a particular interest, as more than 70% of the known antibiotics are produced using bacteria of the *Streptomyces* family. Our ultimate goal consists in identifying σ-factors binding sites upstream of coding parts of the *Streptomyces coelicolor* genomic sequence.

Starting from the last year, we undertook a comparative genomics approach, based on comparison of several phylogenetically related genomes. The method is based on the hypothesis that orthologous proteins of two closely related bacteria must have kept the same regulatory mechanisms, and therefore share common regulatory sites. At the first stage, we run BLASTP [27] on pairs of related bacteria in order to identify orthologous genes between those two organisms. Extracted upstream sequences of those genes were then submitted to MEME [28] program that made a local alignment so as to find motifs shared by the input sequences. Finally, motifs found were then searched for in all *Streptomyces coelicolor* promoter sequences, using the MAST software [29]. Some intermediate results of this approach are reported in [50][51].

In bacteria, binding sites are usually represented by two *boxes* (short sequence motifs) separated by a spacer which can vary slightly among genes. However, MEME is unable to discover a pair of boxes with a variable spacer. This was confirmed by a specific small example of SigR promoter sequences of *Mycobacterium tuberculosis*. Other programmes like BIOPROSPECTOR [47] are able to find variable-spacer motifs, but only on a limited set of sequences and not on the whole-genome level.

This led us to develop, during this year, our own software, to perform a comparative analysis of putatively orthologous genes between *two* genomes. The idea of this approach is to compare upstream regions of pairs of orthologous genes, then to extract all conserved motifs (two boxes with a variable spacer) and then to compare them in order to identify motifs common to several upstream sequence pairs.

An important consideration is that homologous σ-factors in two different bacteria can have slightly different binding sites. This means that detected motifs, shared by pairs of sequences, can be only a part of a "real" motif recognized by σ-factors of respective organisms. Therefore, at the second step, we try to extend each

found motif within the corresponding sequences coming from the same genome. This is done by a recursive procedure that analyses the alignment of all sequences which share a motif, and clusters sequences according to the proportion of the letter found at a given position in the alignment. This procedure results in an extended motif specific to each organism.

Running this method on two bacteria (*Streptomyces coelicolor* and *Mycobacterium tuberculosis*) yielded a high number of candidate motifs (about 15000) which had to be filtered. Two scoring methods have been implemented for this purpose, one using the average of similarity scores between grouped motifs, another by computing the ratio between the number of motifs in upstream regions and the number of motifs in both strands of the whole genome. The higher the ratio is, the more significant the motif is supposed to be. Other approaches for discriminating significant motifs are planned to be applied.

Note that our analysis is based on the comparison of two species – this is because if we use more species, binding sites of a given σ-factor would become too divergent to share a significant common motif. However, it would be still interesting to make a pairwise comparison of several bacteria and to compare results, that we plan to do.

4.1.3. Analysis of multiple repeats in bacteria

This project addresses the issue of repeated DNA sequences, occurring several (more than two) times in a genome. The presence of repeated sequences is a well-known feature of bacterial genomes. Their biological function differs greatly: in one case, a repeat can be about a thousand nucleotides long and contain coding open reading frames (for example, a mobile element); in other cases, a repeat can correspond to a regulatory element located in intergenic regions. Moreover, repeated sequences can be strongly conserved not only within one genome, but also across different (in some cases remotely related) genomes.

There are several programs specifically devoted to the computation of pairwise repeats within a given genomic sequence [45][53][46]. However, there is no method to systematically compute *clusters of repeats*, i.e. sequences that have multiple occurrences in a genome. To compute such clusters, we created a software program REPCLUSTER [21].

The REPCLUSTER software can be used for the analysis of repeated sequences in one genome as well as for a simultaneous analysis of repeats in several genomes. In order to identify multiple repeats in a genome, we first apply the YASS software [10] (see Section 5.3) and find all strong local similarities within the genome, regarded as two-copy repeats. All possible repeated sequences found by YASS are then grouped into clusters, with the goal that each cluster contains all copies of the same repeated biological element. A method of "cores" is used at the clusterization step. Its main idea consists in using most conserved parts of a repeat, called "cores", for controlling the clusterization process. As a result of clusterization, all sequences found by YASS are grouped into clusters, such that each cluster corresponds to an individual repeated element and *vice versa*, all (approximate) copies of the same repeated element belong to the same cluster.

We run our method on the *Neisseria meningitidis* genome and obtained a number of interesting clustered repeats, some of which have a known well-identified biological function. For example, one resulting cluster embraced several hundreds of ρ-independent terminators. Several other clusters corresponded to mobile IS-elements (IS30, IS1016C2, IS1106). Besides of those known elements, some interesting unknown repeats have been detected. For example, we found a cluster of sequences of about 26bp long, which are highly distributed in the genome (a few hundred copies). These repeated sequences form a complex palindromic structure and are located in intergenic regions only, which suggests their possible regulatory role. Moreover, this complex repeated element is often located in regulatory regions of genes involved in bacterial pathogenesis. A similar phenomenon (highly repeated element occurring often upstream of genes involved in pathogenesis) has been also observed in other bacteria. Therefore, these repeats could be interesting objects for further investigation.

4.1.4. Computer analysis of RNA-mediated regulation in bacteria

4.1.4.1. Computer prediction of attenuators and analysis of metabolism of various amino acids in proteobacteria.

We applied bioinformatics and comparative genomics techniques to the prediction of attenuator signals in bacterial genomes, that regulate operons responsible for biosynthesis of several amino acids in gamma- and alpha-proteobacteria, some low-GC Gram-positive bacteria and some other remotedly related bacteria.

This analysis allowed us to identify a large number of candidate attenuators and predict amino acid(s) responsible for the regulation. We also demonstrated a variability of regulatory mechanisms for the amino acid biosynthetic pathways even in closely related bacteria, and allowed for functional annotation of hypothetical genes encoding transporters and enzymes. In particular, three new families of histidine transporters have been predicted, orthologs of yuiF and yvsH of *Bacillus subtilis*, and lysQ of *Lactococcus lactis*.

As another consequence, we showed the diversity and evolutionary lability of regulatory mechanisms based on formation of alternative RNA structures, especially in low-GC Gram-positive bacteria. On the other hand, we demonstrated an ancient origin of this regulatory mechanism. Indeed, we found possible attenuators of amino acid biosynthetic genes not only in proteobacteria, but also in low-GC Gram-positive bacteria, Bacteroidetes/Chlorobi, and, notably, in very ancient Thermotogales and *Deinococcus radiodurans* bacteria. This work is described in publication [13].

4.1.4.2. Prediction of regulatory riboswitches and comparative analysis of methionine regulon in low G/C Gram-positive bacteria

Riboswitches are structures that are formed in mRNA and regulate gene expression in bacteria. Unlike other known RNA regulatory structures, they are directly bounded by small ligands. The mechanism by which gene expression is regulated involves the formation of alternative structures that, in the repressing conformation, causes a premature termination of transcription or inhibition of translation initiation. Riboswitches regulate several metabolic pathways including the biosynthesis of vitamins (e.g. riboflavin, thiamin and cobalamin) and the metabolism of methionine, lysine and purine. Candidate riboswitches have also been observed in archaea and eukaryotes. The taxonomic diversity of genomes containing riboswitches and the diversity of molecular mechanisms of regulation suggest that riboswitches represent one of the oldest regulatory systems [14].

Using the comparative genomics approach, we worked on the regulation mechanism of the methionine biosynthesis and transport genes in bacteria, which is rather diverse and involves two RNA-level regulatory systems and at least three DNA-level systems. Using comparative analysis of genes, operons and regulatory elements, we described methionine regulons in available genomes of Gram-positive bacteria. A large number of methionine-specific RNA elements, S-boxes and T-boxes were identified. These elements have been shown to be widely distributed in Bacillales and Clostridia, whereas methionine-specific T-boxes occurred mostly in Lactobacillales. Positional analysis of methionine-specific regulatory sites complemented by genome context analysis lead to identification of new members of the methionine regulon, both enzymes and transporters. This work has been published in [12].

4.1.5. Genome regulation and DNA curvature

Interactions of geometry with molecular biology is one of the new subjects of our team. As a part of our research on gene regulation, we study the DNA curvature. DNA curvature has been shown to play an important role in a number of biological processes: transcription initiation, DNA replication, etc. The general goal of our work described here is to study the involvement of the DNA curvature in gene regulation. In this section, we describe two pieces of work on DNA curvature that we have done “in parallel” during this year.

4.1.5.1. Computation of the DNA curvature

Various models of DNA curvature have been proposed in the literature but the general idea consists in representing the DNA as a 3-dimensional tube of a constant diameter [31]. There are several software programs for modelling the DNA curvature [49][42], that use different approaches but, for all of them, the computation of the curvature depends on a user-defined parameter that corresponds to the width of the sliding window. Changing the parameter often implies large variations in the obtained results. Using new results of discrete geometry (see Section 6.2), a software program for computing the DNA curvature has been developed in the framework of Franck Rapaport’s DEA work [24]. This program does not require any user-defined parameter and enables to detect all curvature values variations in DNA. First results obtained on some genes, for which promoter regions are well known for their strong curvature values, are encouraging and tests are still in

progress. A summary of this approach was presented in the framework of the LIRIS seminars in Lyon in October and a paper describing this work is in preparation.

We plan to continue this work by first integrating last improvements related to the computation of the discrete curvature of 3D noisy curves [22]. Moreover, we plan to develop a user-friendly interface that would allow biologists to use this software.

4.1.5.2. Computer analysis of DNA curvature of rrn regulatory regions in bacteria.

We used a combination of two techniques - prediction of protein binding sites as well as of DNA curvature - to study FIS/H-NS mediated regulation of rrn operons in proteobacteria. The main difficulty of such kind of analysis is the absence or a degenerative form of DNA-binding sequence motifs for both H-NS and FIS proteins. H-NS has been described to bind non-specifically to DNA and to favor intrinsically curved regions. Based on this knowledge, we used the CURVATURE software [49] in order to predict possible H-NS binding sites upstream of rrn operons in proteobacteria containing H-NS.

Significant conserved DNA curvature signals have been found in rrn promoter regions around -100, -120 positions mostly in three phylogenetic groups: Enterobacteriales, Vibrionales and Pasteurellales. Moreover, FIS binding sites have also been detected only in these three phylogenetic groups. In other proteobacteria, only few significant curved regions have been detected and grouped FIS sites have not been found in rrn promoter regions. This implies a strong correlation between grouped FIS DNA-binding motifs and significant conserved DNA curvature in rrn promoter regions in proteobacteria. As a conclusion, H-NS/FIS mediated antagonistic regulation of rrn operons is possibly evolutionarily conserved only in these three gamma-proteobacterial groups. From a more general perspective, we demonstrated that DNA curvature signals are conserved among various bacteria and a simultaneous prediction of conserved curved DNA sites as well as protein binding sites is important for the evolutionary analysis of gene regulation. To our knowledge, this work is the first attempt of comparative analysis of regulatory sites in bacteria at both DNA sequence and structural levels. A paper describing this work [25] is submitted to a journal.

5. Software

5.1. grappe

Keywords: *DNA sequence*, *motif with jokers*, *multiple motif*, *pattern matching*, *string matching*, *text analysis*.

`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. The software has been registered in *APP (Agence pour la Protection des Programmes)* in 2000, and is distributed in several ways:

- through the Web-page of INRIA free software http://www.inria.fr/valorisation/logiciels/index.fr.html,
- from the page http://www.loria.fr/~kucherov/software/grappe/,
- through the platform *Qualité et Sûreté des Logiciels* http://qsl.loria.fr/ that includes *grappe*.

Note that *grappe* has a special version for processing DNA/RNA sequences that is used in our work on promoter analysis, described in Section 4.1.2.

5.2. mreps

Keywords: *DNA sequence*, *maximal repetition*, *repetition search*, *tandem repeat*.

mreps [44] is a program for computing so-called maximal repetitions in DNA sequences. Maximal repetitions are composed of contiguously repeated fragments that are called *periodicities* in computer science literature and *tandem repeats* in biological literature. The development of *mreps* issued from our theoretical work on an efficient search of all exact maximal repetitions in a text.

Today, version 2.5 of *mreps* is distributed under the GPL license in different ways:

- from its Web page at LORIA http://mreps.loria.fr/
- from the Web page of INRIA free software http://www.inria.fr/valorisation/logiciels/index.fr.html
- from the Web server of the *Collaborative Computational Project 11* http://www.hgmp.mrc.ac.uk/CCP11/index.jsp hosted by the *UK Human Genome Mapping Project Resource Centre*.

mreps can be queried through its Web page, as well as through the BIOWEB server of the Pasteur Institute http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html that provides a web interface to existing popular bioinformatics tools. It is integrated to the *Tandem Repeat Data Base (TRDB)*[2] that is developed by the team of Professor Gary Benson in Boston University.

5.3. YASS

Keywords: *DNA sequences*, *approximate repeats*, *distant repeats*, *local alignment*, *sequence comparison*, *similarity regions*, *spaced seeds*.

We develop YASS – a software for computing similarity regions in genomic sequences (local alignment). The first version has been released in January 2003. YASS accepts input sequences in FASTA/MULTIFASTA format and output results in BLAST-tabular/list/alignment formats, along with the e-value and other statistical parameters.

Comparative tests have been done with similar tools such as BLAST-NCBI, BLAT, BLASTZ and REPuter. They showed that YASS is more sensitive than BLAST, due to the use of a new alignment detection strategy and a possible use of spaced and transition-constrained seeds (see Section 6.1.2). In particular, YASS is more sensitive than BLAST on low-scoring similarities. Moreover, YASS is more sensitive than BLAT too, and one of its advantages over PatternHunter software is the possibility of using transition-constrained seeds, which gives an improvement in sensitivity by 15-20% on coding and/or transition-rich regions. Finally, YASS provides better and less redundant alignments compared to the REPuter software.

YASS is available from

- the INRIA software web page http://www.inria.fr/valorisation/logiciels/vie.fr.html,
- the project URL http://www.loria.fr/projects/YASS/,

YASS can also be queried through a Web server http://yass.loria.fr/interface.php. The algorithm of YASS is discussed in Section 6.1.2 in more details. This year, a new YASS version 1.06 has been released, which includes a gapped alignment post-processing (work done with Steven Corroy during his internship).

[2] http://tandem.bu.edu/trf/trf.html

6. New Results

6.1. Word combinatorics and algorithms on sequences

6.1.1. Repetitions in words

During this year, G. Kucherov work on writing a chapter, jointly with R. Kolpakov, for the third volume *Applied combinatorics of words* of Lothaire series. This chapter, entitled *Periodic structures in words* [9], summarizes, in particular, numerous algorithmic and combinatorial results on repetitions (periodicities) in words, that we obtained for the last five years. The book is now in print by Cambridge University Press.

A paper describing one of the previous results on periodicities in words appeared this year in *Theoretical Computer Science* [7].

6.1.2. Local alignment of DNA sequences

Sequence comparison by *local alignment*, used to detect regions of similarity within the same sequence or between two sequences, remains one of the most commonly used tools in bioinformatics. The work described here and its software implementation in the YASS software (see Section 5.3) aims to improve existing heuristic local alignment methods.

Similar to other heuristic local alignment tools, the method of YASS is based on computing small exact repeats, called *seeds*, that are used as "witnesses" (*hits*) of potential larger similarities. Using each individual seed as a hit would be very inefficient, and therefore in YASS, closely located (including overlapping) seeds are grouped together. The grouping is a key step of the method and is done on the basis of statistical criteria based on the Bernoulli model of the sequence. Some of those criteria are inspired from those used in *Tandem Repeat Finder* [30]. A random walk model is introduced to simulate *indel* events (nucleotide insertions and deletions) appearing between seeds. A coin tossing model (k-order geometric series) gives an upper bound of the maximal accepted distance between seeds found.

Groups of seeds are computed on the fly using a special automaton to update the number of matches or mismatches. Formed groups are then tested by trying to extend them into high-scoring local alignments.

Compared to the popular BLAST software [26] or to the more recent Pattern-Hunter algorithm [48], the YASS method for forming hits is more flexible and adapts to the underlying sequence model. Selectivity/sensitivity ratio is greatly improved over BLAST by using a special additional parameter, called *group size*, equal to the number of matching nucleotides of the group. A further gain in sensitivity is achieved by using *spaced seeds* (see also the next section).

Finally, another innovation introduced by YASS is that it can use so-called *transition-constrained seeds*, that can further improve the sensitivity/selectivity trade-off, especially if the target similarities are well-specified (e.g., restricted to coding regions).

Both improvements of YASS are described in [10][20]. These papers contain also experimental results confirming a better practical performance of YASS over the BLAST program.

6.1.3. Estimation of seed sensitivity

Recently it has been understood that using *spaced seeds* for similarity search is significantly more efficient compared to traditionally used contiguous seeds. On the other hand, multi-seed strategies appeared to be another efficient improvement over the usual single-seed approach. This posed new important questions: how to choose "the best" spaced seed? How to compare different seed-based algorithms?

An answer to these questions requires to specify the notion of a "good alignment" that we want to capture by our search. This, in turn, requires a probabilistic model of those alignments. Different models have been considered in the literature, e.g. Bernoulli model, Markov model or Hidden Markov model. All of them, however, capture *local properties* of alignments but do not allow to specify their *global properties* such as the overall score for example.

6.1.3.1. Homogeneous alignments.

In this context, we proposed a new approach for measuring the sensitivity of a similarity search strategy. The approach is based on the notion of *homogeneous alignment* that captures the type of alignments that are found by virtually all heuristic algorithms and, on the other hand, that are expected to be found.

In collaboration with Yann Ponty (LRI Orsay), we developed a probabilistic model for homogeneous sequences alignments, and proposed an algorithm for random generation of those alignments, and on the other hand, an algorithm for measuring the sensitivity of a seed-based search strategy with respect to those alignments. An important conclusion of this study is that ignoring the property of homogeneity introduces a bias in measurement. This work has been published in [18].

6.1.3.2. General framework and its application to subset seeds.

Recently, we developed a general approach to automatically obtain an efficient algorithm for various instances of the seed sensitivity problem. The approach treats separately three components of the seed sensitivity problem – a set of target alignments, an associated probability distribution, and a seed model – that are specified by distinct finite automata. We showed that once these three components are specified, one can construct, using a single general method, a dynamic programming algorithm for computing seed sensitivity. Several algorithms proposed by other authors [33][32] can be obtained as particular cases of our approach, obtaining the same complexity bounds.

The proposed approach has been applied to a new seed model, called *subset seed*. The interest of subset seeds is that they are more expressive than ordinary spaced seeds, but still allow to be efficiently located using a direct hashing method. Note that subset seeds capture transition-constrained seeds, used in our YASS software (see Section 6.1.2).

We proposed an efficient automaton construction for the set of alignments detected by subset seeds. This automaton is a key component of the algorithm for computing seed sensitivity. Interestingly, instantiated to the case of ordinary spaced seeds, our construction yields an automaton with a number of states smaller than the automaton proposed in previous works.

We also provided experimental evidence to the efficiency of our approach, by performing experimental seed design and testing them on real genomic data. This work, published as an INRIA research report [23], is now submitted to an internatioinal conference.

6.1.4. Approximate pattern matching using multiple seeds

Most of existing approximate pattern matching and local alignment methods are based on the common filtering idea: the algorithm tries to filter out fragments of the text that have no chance to match the pattern. Some well-known such methods (PEX, error PEX) are based on the contiguous text fragments (seeds) that match exactly (or with one error) respective fragments of the pattern. More recently, methods based on spaced seeds have been introduced [34] and have been shown to improve the efficiency.

In this context, we have proposed a new method based on *multiple spaced seeds* that enables a considerable increase of the filter selectivity. The idea here is to combine multiple seeds to solve all the instances of an (m, k) problem (motif of size m with at most k substitution errors). An instance is solved if it contains *one of the input seeds*.

We proposed efficient dynamic programming algorithms to compute different properties of seed families, such as the so-called optimal threshold or the contribution of each seeds of the family. We also proposed several techniques to design efficient lossless seeds and obtained tight asymptotic bounds on the number of jokers of lossless seeds for a given number of substitution errors. Finally, we proposed an heuristic genetic programming algorithm that allows, in combination with above-mentioned dynamic programming algorithms, to design efficient seed families in the general case.

An experimental program has been implemented for designing efficient multiple seeds. The main target application of the program is the design of oligonucleotides for DNA chip experiments. Performed large-scale computational experiment demonstrates that our method can efficiently and exhaustively compute, in a large sequence sample of several billion of letters, all strings uniquely occurring in the sample, up to a given number of substitution errors. Those strings constitute a set of candidate oligonucleotides for DNA chip design.

A paper describing this work has been published in [19]. An extended version is submitted to a journal.

6.2. Discrete geometry

6.2.1. Noisy curves

The recognition of digital objects, such as discrete lines and arcs, is an important topic in discrete geometry that has been subject of numerous works [36][54]. We got interested in the notion of "noisy" digital objects and in their detection. This problem has a direct application in image processing, in particular when existing geometrical shapes have to be interpreted in some images.

6.2.1.1. Blurred segments

We introduced a new concept – *fuzzy* or *blurred segments* – that enables a flexible segmentation of discrete curves by taking into account a noise present in them.

A blurred segment is an 8-connected sequence of points that belong to an arithmetical discrete line with a given thickness. A parameter – the order of a blurred segment – controls the level of the allowed noise via the thickness of the discrete line bounding the blurred segment. Adding a point to a blurred segment amonts to compute the slope and the thickness of the new bounding discrete line. We showed that this computation can be done with a simple method. This led to an incremental and very efficient algorithm for splitting a discrete curve into blurred segments of fixed order.

A paper on this subject was presented to the IWCIA conference in May 2003 [41], and an extended version of this work is accepted to a special issue of Discrete Applied Mathematics [6].

More recently, in collaboration with Fabien Feschet from LLAIC (Clermont-Ferrand), we proposed new results on a restriction in the class of blurred segments in order to guarantee the optimality in the recognition process. This work is now submitted to an international conference [22].

6.2.1.2. Multi-order analysis

A possible application of the above approach occurs in the area of document analysis. In collaboration with Antoine Tabbone and Laurent Wendling of the QGAR team, we designed an algorithm for the polygonal approximation of noisy curves from a multi-order analysis algorithm. Due to the notion of blurred segment, this algorithm does not need to fix parameters and automatically provides a partitioning of a discrete curve into its meaningful parts. This work was presented at the CIFED conference in June 2004 [16] and at the International Conference on Pattern Recognition in August 2004 [17]. An extended version of this work is submitted to the Electronic Letter on Computer Vision and Image Analysis (ELCVIA).

6.2.1.3. Estimation of tangents to a noisy discrete curve

We proposed a new notion of discrete tangent, adapted to 2D noisy curves. It relies on the definition of discrete tangents given by Anne Vialard in 1996 [52], and on the previously mentioned definition of blurred segment together with the corresponding recognition algorithm.

We proposed a variant of the existing algorithm for blurred segment recognition. The new algorithm recognizes, in linear time, a blurred segment of fixed order centered at a given point. An approximate discrete tangent of order d at a point P of a noisy curve is defined to be the longest centered blurred segment of order d recognized by the algorithm at point P. We then obtained an algorithm that computes the parameters of an approximate discrete tangent at each point of a discrete curve. From this algorithm, we can deduce several approximate parameters such as the normal vector or the curvature at each considered point. This work was presented at the *Vision Geometry* conference held in January 2004 in San Jose [15].

During the DEA work of Franck Rapaport [24], we proposed an extension of this technique to discrete 3D noisy curves. An algorithm to compute the curvature at each point of a 3D discrete noisy curve has been obtained. This method was then applied to the computation of the curvature of the DNA molecule (see section 4.1.5). A summary of these results was presented in October in the framework of LIRIS seminars in Lyon and a paper describing this work is in preparation.

6.2.2. Discrete convexity and concavity, polygonal decomposition of discrete sets

The study of the convexity of a discrete region of the plane can be reduced to particular figures called hv-convex polyominoes. Previously, we developed a linear-time incremental algorithm to detect the convexity of such polyominoes [40].

Two years ago, we established contacts with the University of Hamburg, namely with Professor Ulrich Eckart and his student Helene Reiter. In collaboration with Helene Reiter, we developed a linear-time algorithm for decomposition of the boundary of a plane digital object into convex and concave parts. Such a decomposition is very useful for describing the form of an object. The obtained algorithm uses properties of discrete straight lines for the convex case [40], and extends them to the concave case. This work was presented in Dagstuhl in March, during a workshop on *Geometric properties from incomplete data*, and a paper is being published by Kluwer in a book containing selected and reviewed papers of this workshop [11].

6.2.3. Digital plane recognition

A naive digital plane with integer coefficients is defined as a subset of points $(x, y, z) \in \mathbb{Z}^3$ verifying a double inequality $h \leq ax + by + cz < h + \max\{|a|, |b|, |c|\}$, where $(a, b, c, h) \in \mathbb{Z}^4$. Given a finite subset of $\mathbb{Z}^3$, the problem is to determine whether or not there exists a naive digital plane containing it. This question is rather classical in the field of discrete geometry.

With Yan Gerard (LLAIC, Clermont-Ferrand) and Paul Zimmermann (SPACES team), we proposed a new algorithm that solves this problem. The algorithm uses a strategy of optimization in a set of triangular facets (called triangles). The problem consists in finding among a particular set of triangular facets the one which cuts the axis Oz at the highest point. Instead of enumerating their z-coordinates, we suggest an original strategy based on the evaluation of a linear form. A short program code (less than 300 lines) solving the problem is available on the Web[3].

This work was presented at the Days of the *Action Spécifique Algorithmic geometry and discrete geometry* that were held last September. Moreover, a paper describing this work is accepted to *Discrete Applied Mathematics* [8].

7. Other Grants and Activities

7.1. Regional Initiatives

Our team is involved in the *Pôle de Recherche Scientifique et Technologique (PRST) Intelligence Logicielle*, and in particular in the theme *Bioinformatique et Applications à la Génomique* of that project. In this framework, we collaborate with the *Laboratoire de Génétique et de Microbiologie de l'Université Henri Poincaré de Nancy*.

7.2. National Initiatives

Members of the team participate in the *Action Spécifique "Indexation de texte et découverte de motifs" (AS 185)*[4] of the *CNRS*, in the framework of the *Réseau Thématique Pluridisciplinaire Bioinformatique (RTP 41)*. Gregory Kucherov is a co-animator of this working group.

Members of the team also participate in the *Action Spécifique "Géométrie Algorithmique et Géométrie Discrète"*, two meetings of the group were held this year: the first one in January in Lyon and the second one in September in Paris.

We are a part of the project REPEVOL accepted this year within the ACI IMPBio (*Action Incitative Coopérative "Informatique, Mathématiques et Physique pour la Biologie"*) funded by the French government. The project is joint with LIRMM, *Centre d'Ecologie Fonctionnelle et Evolutive* and *Institut de Génétique Humaine* of Montpellier, and Boston University, USA.

[3] http://www.loria.fr/~debled/plane/
[4] http://degas.lirmm.fr/ASIM/

We have additionally an active collaboration with the LLAIC (Clermont-Ferrand) and the LRI (Orsay).

7.3. International Initiatives

Our team participated in an *Arc-en-Ciel* collaborative project with the University of Haifa in Israel on the subject of DNA curvature. This collaboration resulted in publication [25].

We collaborate with the team of Prof. Gary Benson from Boston University within the REPEVOL project of the ACI IMPBio (see previous Section).

Our collaboration with the bioinformatics group of the Warsaw University (J. Tiuryn, A. Gambin) resulted this year in the submission of a french-polish Polonium project (acceptance pending).

Finally, we have an active collaboration with the Institute of Mathematical Problems in Biology in Puschino, Russia (group of M. Roytberg) and the University of Hamburg, Germany (group of Prof. U. Eckart), as witnessed by visits of researchers of those groups to our group (see the next section).

7.4. External visitors

Roman Kolpakov, a post-doc at the University of Liverpool and an old collaborator of ADAGE, visited our group for one week in April 2004 to work with G. Kucherov on a joint book chapter.

Mikhail Roytberg, senior researcher of the Institute of Mathematical Problems in Biology in Puschino (Russia), visited our team for three month (May-July 2004) as an INRIA invited professor.

Mathieu Giraud, a PhD student from the University of Rennes made a three days visit to our team in June 2004, and gave a talk *Des architectures reconfigurables pour accélérer les calculs* at the bioinformatics seminar of LORIA.

Fabien Feschet, *maître de conférences* of the LLAIC, Clermont-Ferrand, came to LORIA for 3 days in July to work on an optimal algorithm of blurred segmentation.

In August 2004, we received, during one week, Ania Gambin and Slawomir Lasota, both lecturers at Warsaw University, as well as former members of their group Radek Szklarczyk (now PhD student in at the Vrije Universiteit, Netherlands) and Rafal Otto (now at CERN, Switzerland). This meeting was an opportunity for us to make mutual presentations of research work on biosequence analysis and to discuss further collaborations.

Helene Reiter, a PhD student from the University of Hamburg (Germany) made a one-week visit to our team in November 2004 to work on the polygonal decomposition of discrete sets.

8. Dissemination

8.1. Services

G. Kucherov served on the program committees of the *5èmes Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)* (Montréal, Canada, June 2004) and the 4th Workshop on Algorithms in Bioinformatics (Bergen, Norway, September 2004).

I. Debled-Rennesson served on the program committees of the International Workshop on Combinatorial Image Analysis (New Zealand, December 2004) and the 12th International conference on Discrete Geometry in Computer Imagery (Poitiers, France, April 2005).

I. Debled-Rennesson is an elected member of the CNU (27th section) and, in the framework of this position, she participated in qualification and promotion sessions for *maîtres de conférences*. She is a member of the IAPR technical committee on discrete geometry (TC18) [5].

G. Kucherov is a member of the *Commission de Spécialistes* of the *Université Henri Poincaré Nancy 1*.

J.-L. Rémy and L. Noé are both members of the *Conseil du Laboratoire* of LORIA.

[5] http://www.cb.uu.se/~tc18/

8.2. Teaching

I. Debled-Rennesson supervised the DEA training of Franck Rapaport in February-August 2004, and the internship of Trung Nguyen (magistère ENS, Paris) in June-August 2004, as well as a research project of an ESIAL student.

G. Kucherov et L. Noé supervised the internship of Steven Corroy (*Ecole Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble*) in July-August 2004.

Jointly with D. Kratsch (*Université de Metz*), G. Kucherov taught the course *Algorithmics of discrete structures* of *DEA d'Informatique* of Nancy (specialization *Algorithmique Numérique et Symbolique*). He also delivered lectures on bioinformatics to the DESS *Ressources Génomiques et Traitements Informatiques* of the *Université Henri Poincaré de Nancy*, and to the bioinformatics program of the *École des Mines de Nancy*.

In the framework of their *monitorat*, L. Noé and F. Touzain delivered programming courses at the *Université Henri Poincaré de Nancy* (MIAS2, IUT de Metz).

8.3. Participation in meetings, seminars, invited talks

8.3.1. Meetings, tutorials, conferences, invited seminar talks

During this year, I. Debled-Rennesson made the following talks:

- at the *Vision Geometry XII Conference, SPIE*, San Jose, USA in January 2004,
- at the workshop on *Geometric properties from incomplete data*, Dagstul (Germany) in March 2004,
- at LIRIS (Lyon) seminar, in October 2004,
- at the workshop *Contenu Informatif des Images Numériques* (ENS Cachan) in November 2004,

She presented a poster, jointly with two members of the QGAR team, at the ICPR conference in Cambridge, United Kingdom, in August 2004.

She also participated in two meetings of the *Action Spécifique Géométrie Algorithmique et Géométrie Discrète*, with Jocelyne Rouyer in January 2004 (LIRIS, Lyon) and in September 2004 (ESIEE, Paris).

L. Noé gave a talk in December 2003 at the seminar *Mathématiques pour le Génome* in Evry.

G. Kucherov gave a talk in December 2003 at the meeting of the IMPG working group on word statistics that took place at LRI, Orsay.

G. Kucherov and L. Noé were invited in Mars 2004 to give a talk at the seminar of the *SYMBIOSE* project-team in Rennes.

G. Kucherov and L. Noé participated in the BIBE conference in Taichung (Taiwan) in May 2004. L. Noé gave a talk at that conference.

G. Kucherov and L. Noé participated in the meeting of the *Action Spécifique Indexation de texte et découverte de motifs* in Nantes held on 27-28 May 2004. L. Noé made a talk there on multi-seed filtering techniques.

L. Noé gave a talk at the JOBIM conference in Montréal (Canada) in June 2004.

G. Kucherov, L. Noé and M. Roytberg participated in the CPM conference in Istanbul (Turkey) in July 2004. G. Kucherov gave a talk at that conference.

G. Kucherov, A. Vitreschak and M. Roytberg participated in the 4th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS) that was held in Novosibirsk, Russia. G. Kucherov and M. Roytberg gave both a talk at this conference and A. Vitreschak presented a poster.

G. Kucherov, A. Vitreschak, L. Noé and F. Touzain are going to participate in the meeting of the *Action Spécifique Indexation de texte et découverte de motifs* in Lille on 9-10 December 2004.

8.3.2. Visits of team members

In December 2003 and in June 2004, J.-L. Rémy visited Y. Pétermann at the University of Genève, for a total time of 3 weeks.

8.4. Participation in juries

I. Debled-Rennesson participated in the jury of PhD thesis of X. Hilaire (LORIA) in January 2004. She was also a reviewer of Helene Reiter-Dorksen's PhD thesis (Hamburg University, Germany) defended in November 2004.

G. Kucherov was a reviewer of the PhD thesis of Johann Pelfrêne, defended in June 2004 at the University of Rouen.

9. Bibliography

Major publications by the team in recent years

[1] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Detection of the Discrete Convexity of Polyominoes*, in "Discrete Applied Mathematics", vol. 125, n° 1, Jan 2003, p. 115-133.

[2] R. KOLPAKOV, G. BANA, G. KUCHEROV. *mreps : efficient and flexible detection of tandem repeats in DNA*, in "Nucleic Acids Research", vol. 31, n° 13, Jul 2003, p. 3672-3678.

[3] R. KOLPAKOV, G. KUCHEROV. *Finding approximate repetitions under Hamming distance*, in "Theoretical Computer Science", vol. 303, n° 1, June 2003, p. 135-156.

[4] R. KOLPAKOV, G. KUCHEROV. *Finding Maximal Repetitions in a Word in Linear Time*, in "Proceedings of the 1999 Symposium on Foundations of Computer Science, New York, U.S.A.", IEEE Computer Society, October 1999, p. 596–604.

[5] G. KUCHEROV, P. OCHEM, M. RAO. *How many square occurrences must a binary sequence contain?*, in "The Electronic Journal of Combinatorics", vol. 10, n° 1, January 2003.

Articles in referred journals and book chapters

[6] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Linear Segmentation of Discrete Curves into Fuzzy Segments*, in "Discrete Applied Mathematics", 2004.

[7] J.-P. DUVAL, R. KOLPAKOV, G. KUCHEROV, T. LECROQ, A. LEFEBVRE. *Linear-time computation of local periods*, in "Theoretical Computer Science", vol. 326, n° 1–3, October 2004, p. 229–240.

[8] Y. GÉRARD, I. DEBLED-RENNESSON, P. ZIMMERMANN. *An elementary digital plane recognition algorithm*, in "Discrete Applied Mathematics", 2004.

[9] R. KOLPAKOV, G. KUCHEROV. *Identification of periodic structures in words*, in "Applied combinatorics on words", J. BERSTEL, D. PERRIN (editors)., Lothaire books, Cambridge University Press, 2004.

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

[11] H. REITER-DORKSEN, I. DEBLED-RENNESSON. *Convex and Concave Parts of Digital Curves*, in "Geometric Properties from Incomplete Data", Kluwer, December 2004.

[12] D. RODIONOV, A. VITRESCHAK, A. MIRONOV, M. GELFAND. *Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems*, in "Nucleic Acids Research", vol. 32, n° 11, June 2004, p. 3340-3353.

[13] A. VITRESCHAK, E. LYUBETSKAYA, M. SHIRSHIN, M. GELFAND, V. LYUBETSKY. *Attenuation regulation of amino acid biosynthetic operons in proteobacteria : comparative genomics analysis*, in "FEMS Microbiology letters", vol. 234, n° 2, May 2004, p. 357-370.

[14] A. VITRESCHAK, D. RODIONOV, A. MIRONOV, M. GELFAND. *Riboswitches: the oldest mechanism for the regulation of gene expression?*, in "Trends in Genetics", vol. 20, n° 1, January 2004, p. 44-50.

Publications in Conferences and Workshops

[15] I. DEBLED-RENNESSON. *Estimation of Tangents to a Noisy Discrete Curve*, in "Vision Geometry XII, Electronic Imaging, San Jose, California, USA", L. J. LATECKI, D. M. MOUNT, A. Y. WU (editors)., Proceedings of the SPIE, vol. 5300, January 2004, p. 117-126.

[16] I. DEBLED-RENNESSON, S. TABBONE, L. WENDLING. *Approximation polygonale à partir d'une analyse multi-ordres des points de contour*, in "Colloque International Francophone sur l'Ecrit et le Document - CIFED'2004, La Rochelle, France", June 2004.

[17] I. DEBLED-RENNESSON, S. TABBONE, L. WENDLING. *Fast polygonal approximation of digital curves*, in "17th International Conference on Pattern Recognition - ICPR'04, Cambridge, United Kingdom", vol. 1, August 2004, p. 465-468.

[18] G. KUCHEROV, L. NOÉ, Y. PONTY. *Estimating seed sensitivity on homogeneous alignments*, in "Proceedings of the IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE2004), May 19-21, 2004, Taichung (Taiwan)", IEEE Computer Society Press, April 2004, p. 387–394.

[19] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *Multi-seed lossless filtration*, in "Proceedings of the 15th Annual Combinatorial Pattern Matching Symposium (CPM), July 5-7, 2004, Istanbul (Turkey)", S. SAHINALP, S. MUTHUKRISHNAN, U. DOGRUSOZ (editors)., Lecture Notes in Computer Science, vol. 3109, Springer Verlag, 2004, p. 297–310.

[20] L. NOÉ, G. KUCHEROV. *Improved hit criteria for DNA local alignment*, in "Proceedings of the 5th Open Days in Biology, Computer Science and Mathematics (JOBIM), June 28-30, 2004, Montréal (Canada)", June 2004.

[21] A. VITRESCHAK, L. NOÉ, G. KUCHEROV. *Computer Analysis of Multiple Repeats in Bacteria*, in "Proceedings of the 4th International conference on Bioinformatics of Genome Regulation and Structure (BGRS), July 25-30, 2004, Novosibirsk, (Russia)", vol. 2, Institute of Cytology and Genetics, July 2004, p. 297-299.

Internal Reports

[22] I. DEBLED-RENNESSON, F. FESCHET, J. ROUYER-DEGLI. *Optimal Blurred Segments Decomposition*

in Linear Time, Rapport de recherche, n° RR-5334, INRIA, November 2004, http://www.inria.fr/rrrt/rr-5334.html.

[23] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *A unifying framework for seed sensitivity and its application to subset seeds*, Rapport de recherche, n° RR-5374, INRIA, Nov 2004, http://www.inria.fr/rrrt/rr-5374.html.

[24] F. RAPAPORT. *Calcul du rayon de courbure d'une séquence d'ADN*, Stage de DEA, June 2004, http://www.loria.fr/publications/2004/A04-R-276/A04-R-276.ps.

[25] A. VITRESCHAK, S. HOSID, A. BOLSHOY, G. KUCHEROV. *Computer analysis of rrn regulatory signals on both DNA sequence and structural levels in proteobacteria*, Rapport de recherche, n° A04-R-332, LORIA, October 2004.

Bibliography in notes

[26] S. ALTSCHUL, T. 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, n° 17, 1997, p. 3389–3402.

[27] S. ALTSCHUL, T. MADDEN, A. SCHAFFER, 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.

[28] T. BAILEY, C. ELKAN. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*, in "Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology, Menlo Park, California", AAAI Press, 1994, p. 28-36.

[29] T. BAILEY, M. GRIBSKOV. *Combining evidence using p-values: application to sequence homology searches*, in "Bioinformatics", vol. 14, 1998, p. 48-54.

[30] G. BENSON. *Tandem repeats finder: a program to analyse DNA sequences*, in "Nucleic Acids Research", vol. 27, n° 2, 1999, p. 573–580.

[31] A. BOLSHOY, P. MCNAMARA, P. HARRINGTON, E. TRIFONOV. *Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles*, in "Proc. Natl. Acad. Sci. USA", vol. 88, 1991, p. 2312-6.

[32] B. BREJOVA, D. BROWN, T. VINAR. *Vector seeds: an extension to spaced seeds allows substantial improvements in sensitivity and specificity*, in "Proceedings of the 3rd International Workshop in Algorithms in Bioinformatics (WABI), Budapest (Hungary)", R. P. G. BENSON (editor)., Lecture Notes in Computer Science, vol. 2812, Springer, September 2003.

[33] J. BUHLER, U. KEICH, Y. SUN. *Designing seeds for similarity search in genomic DNA*, in "Proceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB03), Berlin (Germany)", ACM Press, April 2003, p. 67-75.

[34] S. BURKHARDT, J. KÄRKKÄINEN. *Better filtering with gapped q-grams*, in "Fundamenta Informaticae", vol. 56, n° 1-2, 2003, p. 51-70.

[35] J.-M. CHASSERY, A. MONTANVERT. *Géométrie discrète en imagerie*, Hermès, Paris, 1991.

[36] D. COEURJOLLY, L. TOUGNE, Y. GÉRARD, J.-P. REVEILLÈS. *An Elementary Algorithm for Digital Arc Segmentation*, in "Electronic Notes in Theoretical Computer Science", vol. 46, 2001.

[37] M. CROCHEMORE, C. HANCART, T. LECROQ. *Algorithmique du texte*, Vuibert Informatique, 2001.

[38] M. CROCHEMORE, W. RYTTER. *Jewels of Stringology*, World Scientific, 2002.

[39] M. CROCHEMORE, W. RYTTER. *Text algorithms*, Oxford University Press, 1994.

[40] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Detection of the Discrete Convexity of Polyominoes*, in "Discrete Applied Mathematics", vol. 125, n° 1, Jan 2003, p. 115-133.

[41] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Segmentation of Discrete Curves into Fuzzy Segments*, in "Proceedings of the 9th International Workshop on Combinatorial Image Analysis (IWCIA'2003), Palermo, Italy", Electronic Notes in Discrete Mathematics, vol. 12, May 2003.

[42] M. DLAKIC, R. HARRINGTON. *DIAMOD: display and modeling of DNA bending*, in "Bioinformatics", vol. 14, 1998, p. 326-331.

[43] D. GUSFIELD. *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997.

[44] R. KOLPAKOV, G. BANA, G. KUCHEROV. *mreps : efficient and flexible detection of tandem repeats in DNA*, in "Nucleic Acids Research", vol. 31, n° 13, Jul 2003, p. 3672-3678.

[45] S. KURTZ, J. CHOUDHURI, E. OHLEBUSCH, C. SCHLEIERMACHER, J. STOYE, R. GIEGERICH. *REPuter: the manifold applications of repeat analysis on a genomic scale*, in "Nucleic Acids Res.", vol. 29, n° 22, Nov 15 2001, p. 4633-42.

[46] A. LEFEBVRE, T. LECROQ, H. DAUCHEL, J. ALEXANDRE. *FORRepeats: detects repeats on entire chromosomes and between genomes*, in "Bioinformatics", vol. 19, n° 3, Feb 12 2003, p. 319-26.

[47] X. LIU, D. BUTLAG, J. LUI. *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes*, in "Pacific Symposium on Biocomputing", vol. 6, 2001, p. 127-138.

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

[49] E. SHPIGELMAN, E. TRIFONOV, A. BOLSHOY. *CURVATURE: software for the analysis of curved DNA*, in "Comput Appl Biosci", vol. 9, n° 4, 1993, p. 435-40.

[50] F. TOUZAIN, P. LAVIGNE, I. DEBLED-RENNESSON, B. AIGLE, P. LEBLOND, G. KUCHEROV. *Identification of Transcription Factor Binding Sites in Streptomyces coelicolor A3(2) by Phylogenetic Comparison*, in "European Conference for Computer Biology - ECCB'2003, Paris, France", Poster, Sep 2003.

[51] F. TOUZAIN. *Recherche des sites de régulation de la transcription chez Streptomyces Coelicolor A3(2)*, Stage de DEA, Sep 2003.

[52] A. VIALARD. *Geometrical Parameters Extraction from Discrete Paths*, in "6th DGCI", Lecture Notes in Computer Science, vol. 1176, Springer-Verlag, 1996, p. 24-35.

[53] A. VINCENS, C. ANDRÉ, S. HAZOUT. *D-ASSIRC: distributed program for finding sequence similarities in genomes*, in "Bioinformatics", vol. 18, n^o 3, March 2002, p. 446-51.

[54] W. WAN, J. A. VENTURA. *Segmentation of Planar Curves into Straight-Line Segments and Elliptical Arcs*, in "Graphical Models and Image Processing", vol. 59, 1997, p. 484–494.