INRIA

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 2B

# 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**
Hélène Zganic [TR INRIA, 1/4 of the full time]

**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]

**Faculty members**
Jocelyne Rouyer [Maître de conférences, UHP]

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

**Junior technical staff**
Ghizlane Bana [INRIA, till September 30, 2003]

**Post-doctoral fellow**
Alexey Vitreschak [INRIA, from September 1, 2003]

**Internships**
Ougas Elmi Houssein [DESS IDC of Nancy, June-July 2003]
Patricia Lavigne [DESS bioinformatique de Rouen, till July 2003]
Arya Pranjal [Indian Institute of Technology, Delhi, May-July 2003]

# 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 combinatorial 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 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 for our models and algorithms is particularly important for 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 in experimental software systems and to making them available to scientific community. Two 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**, has been developed earlier.

# 3. Scientific Foundations

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

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.

We now briefly describe each of our research areas.

*3.1.1. Text algorithms*

The area of string algorithms (also called text or sequence algorithms) has been very actively developed during the last decade, as witnessed by the publication of several monographs [38][41][36][37]. 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.

String algorithms are also very important from a theoretical perspective. The core of this theory is composed of several algorithms and data structures that undoubtedly make part of "jewels of algorithmics". The best known one is the algorithm of Knuth-Morris-Pratt that is presented of most algorithmics manuals and that has been (and still is) used in many other problems and, on the other hand, has been subject of an interesting and non-trivial mathematical analysis [42]. Other string algorithms play a fundamental role, like the dynamic programming algorithm of computing the longest common subsequence of two sequences that has numerous applications in different areas (like the UNIX `diff` tool, or the well-known Smith-Waterman algorithm for *local alignment* of biological sequences, see Section 6.1.3). Among other algorithms, possibly less known but still very elegant, we can point out the algorithm of checking in linear time the square-freeness of the word [35] or the algorithm of computing in linear time all the palindromes of the word [44].

Besides of this, string algorithms gave rise to very powerful data structures, such as *suffix tree* and the DAWG (*Directed Acyclic Word Graph*). The primary goal of those structures is to perform the text indexation, i.e. to provide a representation of the text that allows to efficiently execute different queries. Moreover, building such a data structure out of the text is done very efficiently too, namely in linear time and in the on-line fashion. We refer to the monograph [41] that is largely devoted to the suffix tree and its different applications.

An important aspect of string algorithms is that they are essentially based on combinatorial properties of words. Many algorithms use word combinatorics to improve their efficiency. That is why the combinatorics on words plays for us an important role and is a subject of our studies too.

To summarize, our goal 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*

Among discrete structures that we study, a special attention is given to discrete sets of the plane or of the space. This research direction, called *discrete* (or *digital*) *geometry* , appeared in the 70's. 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 along these lines [32]:

- a topological viewpoint that is concerned with, for instance, a discrete equivalent of the Jordan's theorem (any simple closed curve partitions the plane into two domains: "inside" and "outside" of the curve),
- a morphological viewpoint that studies the transformations of shapes (morphological analysis),
- an arithmetical viewpoint, introduced by Jean-Pierre Reveillès in 1989 [46], that gives a comprehensive definition to discrete lines and planes.

In our research, we follow the third approach. A discrete straight line on a plane is then a set of points $(x, y)$ of $\mathbb{Z}^2$ verifying the double inequality $\mu \leq ax+by < \mu+\omega$, with $a, b, \mu, \omega$ integers. Properties of discrete lines defined in this way are in close relationship with the properties of integer numbers and therefore are related to the combinatorics on words (cf. the relation to Sturm words for example). Discrete planes in a 3D space are defined in a similar way.

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.

Numerous results based on this approach have been obtained during the last decade. They can be organized into several topics:

- discrete transformations: nearly affine applications, rotations, filters,
- visualization, using properties of discrete objects, e.g. optimal thickness of the objects,
- definition and study of new classes of discrete objects (3D lines, hyperplanes, circles, spheres, simplexes, ...),
- analytical recognition that allows not only to test whether a sequence of points is a line segment or not, but also to give the coefficients of the corresponding analytical inequalities,
- analytical reconstruction enabling a transition between comprehensive representations in discrete and continuous settings,

In our works we are mainly concerned with the last three topics of this list.

*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). Here we describe one of the situations when the probabilistic analysis comes into play.

A computational treatment of genomic data allows to predict some events, for instance on the sequence level (like sequence similarities or repetitions). The biological meaning of those events is not known in general and cannot be determined without a biological expertise. Since the quantity of those data is usually large, we then need a criterion to filter out *a priori* as much of data as possible. To do that, we try to estimate the likelihood of the event, under the hypothesis that the event does not have a biological meaning but occurs completely at random. This hypothesis, called the *null hypothesis*, is widely used in bioinformatics to classify the predicted data. The underlying idea of the approach is that if an event is likely to be "random", then it is unlikely that it represents a biological phenomenon. *A contrario*, only "surprising" events can be biologically significant. A classical example [45] is the unusual abundance, in the *E.coli* genome, of the word `gctggtgg` that has a known biological function.

To implement such a computational procedure one needs to define a probabilistic model of the phenomenon of interest and to perform a corresponding probabilistic analysis. An additional importance of this analysis is that it allows to define a normalized estimator of the "value" of each event and thus to compare different events

which would be difficult to compare otherwise. A well-known example is the Karlin model used to estimate the significance (so-called *p-value*) of local similarities found in two sequences (see Section 6.1.3).

Another common situation when the probabilistic analysis plays an essential role is motivated by algorithmic reasons. The quantity of genomic data and the complexity of problems to solve are such that exact combinatorial algorithms become infeasible more and more often. A usual approach is then to give up the exhaustiveness and to try to build an efficient algorithm that finds *most* of the solutions (instead of finding all of them). However, to justify such an approach one needs to insure that the fraction of results missed by the algorithm is *typically* small. This leads to a probabilistic analysis that demonstrates, with respect to some *probabilistic model*, that, with high probability, the fraction of missed results is small.

The two situations above illustrate the role that the probabilistic analysis play in our research work.

# 4. Application Domains

## 4.1. Bioinformatics

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

### *4.1.1. Introduction*

Discrete models come up in virtually all application areas but one of them plays for us a particular role: this area is molecular biology that studies biological macromolecules – DNA, RNA and proteins. Bioinformatics gives us a source of problems to study and, on the other hand, an excellent field to apply and to test our ideas and methods. The relevance of discrete models is directly explained by the linear structure of biological macromolecules, which naturally leads to their representation as texts over an alphabet of small size. This linear model does not capture *all* properties of biological molecules but most of them, and therefore generally we restrict ourselves to this model. In other words, we are interested in "fingerprints" of biological phenomena in genomic 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. Most of them focus on the sequence level and try to apply our knowledge of sequence analysis methods, gained in our theoretical studies. However, one of those projects, described in Section 4.1.7, goes beyond a pure sequence analysis and tries 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 possibly to bring out 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 seventy percent of the known antibiotics are produced using bacteria of the *Streptomyces* family. Our ultimate goal consists in identifying binding sites of $\sigma$-factors in upstream coding parts of the *Streptomyces coelicolor* genomic sequence.

In our previous work, we attempted to identify transcription factor binding sites using the idea of their statistical "over-representation". The latter was estimated using R'MES software, that identifies DNA motifs with "unexpected" occurrence rate (according to a Markov model of order $m$). Whereas experimentally confirmed promoters have been found within this approach, a big amount of false positives did not allow

us to use this criterion alone to determine *bona fide* regulatory sites. To better understand the "behavior" of regulatory sites, we compared the distribution of some known motifs in promoter and coding regions. For some of the motifs, we did not observe a stronger representation in promoter regions with respect to coding regions. This implied that criteria other than the statistical representation have to be taken into account. All results obtained at that stage were reported in [18].

Starting from this year, we have undertaken a comparative approach based on comparison of genomes of several phylogenetically related bacteria. The method is based on the hypothesis that 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 [25] on pairs of related bacteria in order to identify orthologous genes between those two organisms. We then processed BLASTP results and obtained all combinations of putative orthologous genes between two, three or four species. For each such combination, we extracted upstream sequences of corresponding genes of each organism.

Each set of upstream regions of putative orthologous genes was then submitted to MEME [26] program that made a local alignment so as to find motifs shared by the input sequences. Motifs found were then searched for in all *Streptomyces coelicolor* promoter sequences, using MAST software [27]. The purpose of this search was to possibly generalize the motif and to adapt it to the *Streptomyces coelicolor* genome. Finally, computed motifs were displayed in a graphical form using MAKELOGO[1] program. The whole approach was implemented in a set of PERL programs. Preliminary results of this work were presented in [19][24].

Although the work is still in progress, some interesting results have been obtained with two species (*Streptomyces coelicolor* and *Streptomyces avermitilis*) and three species (*Streptomyces coelicolor*, *Streptomyces avermitilis* and *Mycobacterium leprae*). In the first case, more than 16 "orthology" relationships yielded the same consensual motif present in different sequences. It is all the more interesting as the found motif contains one of experimentally confirmed motifs for some genes. In the second case (three species), a motif has been found that displays a similar structure. These results remain to be experimentally confirmed, but they seem promising as the sensitivity of the approach can be improved by adjusting background models and using additional biological data. Also, information of more species has to be taken into account.

### *4.1.3. Analysis of repeated sequences in proteobacteria genomes*

Another bioinformatics project that we undertook this year addresses the issue of repeated DNA sequences, occurring several (more than two) times in a genome. From the computer science viewpoint, this project is based on the YASS software (see Section 5.3), that can be used to identify *pairs* of repeated sequences in a given genome. On top of YASS, we created another software for computing *clusters* of repeated sequences. In order to exclude parologous genes, we restricted our search to intergenic regions only. Note here that designing a good method for computing multiple repeats out of a set of pairwise repeats is an interesting computer science problem on its own, that amounts, in a certain sense, to compute "quasi-cliques" in a graph. Although we are working on designing such a method, here we concentrate on the bioinformatics aspect of this work only.

The above approach has been applied to identify clusters of repeated sequences in proteobacteria genomes, in particular in *Neisseria meningitidis serotype A and B*. Obtained clusters were then analyzed using various bioinformatics techniques. In one case, we found mobile IS-elements from IS30, IS1016C2 and IS1106 families that are known to be present in proteobacteria genomes. These elements are about 1000 bp long and encode a single protein, transposase, which is a main protein for genetic mobility of IS-elements. In other cases, we found various repeat sequences, that are about 120 bp long and are highly distributed in the genome (40-50 copies per genome). These repeat sequences have a complex structure: inverted repeat at the ends of repeat sequence (about 20 bp long), and a number of short inverted repeats within the repeat sequence. In particular, we focused on one of those repeated sequences and found out that it has an inverted repeat at the ends (20 bp. long), which, in turn, includes another inverted repeat 8 bp. long. Moreover, inside this element we found a tandem repeat of 7 bp. long, present at least four times. We conjecture that this repeated elements

[1] http://www.lecb.ncifcrf.gov/~toms/delila/makelogo.html

could be an RNA gene with a strong secondary structure and could be a new kind of short mobile element (because of its high presence in the genome) and play a biological role in the cell.

We observed that found repetitive elements are often located closely to genes involved in the pathogenesis of the bacterium. Thus, these repeats could be interesting objects for further investigation. That is why we plan to continue this study.

### 4.1.4. Recombination hot spots in human genome

Some modifications of the genome are caused by sequence rearrangements during the recombination. Deletions and duplications could result from abnormal recombinations between sequences that are similar but not homologous. This mechanism deletes a region of one chromosome and inserts it into another chromosome, as illustrated by the Charcot-Marie-Tooth disease that is caused by a chromosomic duplication near the gene of so-called perepherical myelin protein (PMP-22). Others rearrangements could result from homologous recombinations that frequently occur in regions with tandem repeats, due to a strong homology between repeated units. It is therefore very important to determine the location of those regions, called *hot spots*, in order to understand recombination mechanisms and to detect regions preferentially concerned by this rearrangement mechanism.

In collaboration with Marie-Dominique Devignes (team ORPAILLEUR), we undertook a study of hot-spots of human chromosome 22. This work is based on the delimitation of hot spot zones obtained by the laboratory of *Génétique Moléculaire et Biologie du Développement* at Villejuif [33], based on genetics and hybrid radiation maps. **mreps** software (Section 5.2) was used to analyze those zones in order to establish tandem repeat profiles. A profile is defined by the distribution and features of tandem repeats occurring in the sequence. The goal of this approach is to try to establish a correlation between the tandem repeat profile of a sequence and its recombination properties.

Our studies did reveal a correlation between the tandem repeats profile and the location of the sequence with respect to hot spot zones. On the other hand, it turned out to be difficult to generalize this correlation to other chromosomes. Indeed, the relationship between tandem repeats and hot spots is subject to many environment conditions that should be taken into account. This requires further studies that we plan to pursue in future.

### 4.1.5. Computer analysis of lysine-specific RNA regulatory elements in bacteria

We applied bioinformatics and comparative genomics techniques to the analysis of bacterial lysine metabolism and transport genes. Identification of an mRNA structure in the regulatory regions of lysine biosynthetic and transport genes in bacteria allowed us to describe a previously uncharacterized lysine regulon. We named this new highly conserved regulatory mRNA structure the LYS element. By analogy to the previously described metabolite-specific RNA elements, we propose a possible mechanism of the lysine regulation mediated by the LYS element. Using a combination of the analysis of regulatory elements and the genome context analysis, several new LYS-element-regulated transporters, which are possibly specific for lysine, were predicted in various bacteria. We also identified a number of new candidate enzymes from the lysine biosynthetic pathway [13]. Recently, it has been demonstrated by in vitro experiment that lysine can directly regulate the *B. subtilis* lysC gene using termination/antitermination mechanism [40]. Thus, this result may be considered as a confirmation of our hypothesis of lysine riboswitch. From the practical point of view, this work, in addition to our previous analyses of the vitamin-specific regulons [14], gives one more example of the power of comparative genomics for the prediction of regulation and functional gene annotation. Comparative analysis of pathway-specific regulatory sites in bacterial genomes is very effective in this respect [12]. Combination of genomic techniques allowed us to identify candidates for previously missing lysine biosynthetic and transport genes in a variety of bacterial species as well as the riboswitch mechanism of regulation.

### 4.1.6. Computer analysis of transcription attenuators in bacteria

Using comparative genomics approach, we worked on the prediction of candidate transcription attenuators that regulate operons responsible for biosynthesis of branched amino acids, histidine, threonine, tryptophan, and phenylalanine in gamma- and alpha-proteobacteria, low-GC Gram-positive bacteria, Thermotogales and Bacteroidetes/Chlorobi. This analysis allowed us to identify a large number of candidate attenuators and

to predict the amino acid responsible for the regulation. Moreover, this allowed us to show the variability of regulatory mechanisms for the amino acid biosynthetic pathways even in closely related genomes, and to construct a functional annotation of hypothetical genes encoding putative transporters and enzymes. In particular, orthologs of *ygeA* of *E. coli* were predicted to encode branched chain amino acid racemase. Three new families of histidine transporters were predicted: orthologs of yuiF and yvsH of *B. subtilis*, and lysQ of *L. lactis*. In *Pasteurellales*, a single bi-functional aspartate kinase/homoserine dehydrogenase gene thrA was shown to be regulated not only by threonine and isoleucine, but also by methionine. Candidate attenuators were found in some taxonomic groups where such mechanism of regulation was rarely studied (alpha-proteobacteria,) or not studied at all (low-GC Gram-positive bacteria, Bacteroidetes/Chlorobi group and Thermotogales). This analysis, as well as other comparative studies, demonstrates the diversity and evolutionary lability of regulatory mechanisms based on formation of alternative RNA structures, especially in low-GC Gram-positive bacteria. Indeed, we observed candidate histidine attenuators regulating *his* operons in some bacilli and clostridiae. However, in streptococci this operon is regulated by the T-boxe mechanism. This situation is similar to the one with the methionine biosynthesis pathway which is regulated by T-boxes in streptococci, S-box riboswitches in bacilli and clostridiae, and by transcription repression in lactobacilli (paper in preparation).

#### *4.1.7. Genome regulation and DNA curvature*

Here, our general goal was to study, using computer methods, the influence of the DNA curvature on the efficiency of binding sites of some proteins. In particular, we focused on the gene regulation possibly mediated by global regulatory proteins H-NS/FIS in bacteria. The main difficulty of such kind of analysis is the absence or a degenerative form of conserved DNA-binding sequence motifs for both H-NS and FIS proteins. H-NS is described to bind non-specifically to DNA and prefers intrinsically curved regions. Based on this knowledge, we used CURVATURE [47] software in order to predict possible H-NS binding sites upstream of rrn operons in proteobacteria containing H-NS. rrnB operon in Escherichia coli is known to be regulated by FIS/H-NS and regulatory region of this operon contain intrinsically curved region. We analyzed regulatory regions of other six rrn operons in this bacteria and found a curved region, with the center located approximately at -90 - 110 relative to the transcription start site. Analysis of regulatory regions of predicted rrn operons in other proteobacteria showed a high degree of DNA curvature upstream of transcription start site of rrn operons, although the position of the center of curvature in various bacteria differs. This work is a first step towards a general analysis of protein binding sites, using DNA curvature information.

# 5. Software

## 5.1. grappe

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

`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 de Protection de 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

**Key words:** *DNA sequence*, *repetition search*, *maximal repetition*, *tandem repeat*.

**mreps** 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**, started four years ago, issued from our theoretical work on an efficient search of all exact maximal repetitions in a text.

Since that time, **mreps** software has been improved in different ways and applied to several genomic studies. One of them is described in Section 4.1.4. A great deal of work on the development of **mreps** program was done during year 2002. This work resulted in version 2.5 of **mreps**, available now. This version is described in [9] appeared this year in *Nucleic Acid Research*.

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

- from its Web page at LORIA http://www.loria.fr/mreps/
- 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, now in Boston University.

## 5.3. YASS

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

We developed YASS – a software for computing similarity regions in genomic sequences (local alignment). The first version has been released in January 2003. The current version is 1.04.

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. The software is written in ANSI C language. It has been tested under Unix and Windows environments on different genomic sequences (chromosomes of *Saccharomyces cerevisiae*, several *Neisseiria meningitidis* strain sequences, several plant sequences).

Comparative tests have been done with other 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 6.1.3). In particular, YASS is more sensitive than BLAST on low-scoring similarities. On the other hand, YASS output 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 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 developed this summer. The algorithm of YASS is discussed in Section 6.1.3 in more details.

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

# 6. New Results

## 6.1. Word combinatorics and algorithms on sequences

### *6.1.1. Combinatorics of repetitions in words*

The paper that describes our work on the minimal number of square occurrences in binary words appeared this year in the *Electronic Journal of Combinatorics* [11]. This work, joint with Pascal Ochem (PhD student at *LaBRI*, Bordeaux) and Michaël Rao (PhD student at *LITA*, Metz), is devoted to the classical subject of word combinatorics – properties of infinite words with constraints on occurring repetitions. Note that in 1997-1998 we have already done some work on closely related topic[6].

To briefly describe the results obtained in [11], note first that every binary word of length at least 4 contains a *square* (subword $uu$). A. Fraenkel and J. Simpson [39] showed that there exists an infinite binary word containing three *distinct* squares (e.g. 00, 11 and 0101), and this number is minimal. We studied a complementary question of the minimal number of *square occurrences* contained in infinite binary words. We showed that the minimal number of square occurrences contained in binary words of length $n$ is a constant fraction of $n$ when $n$ goes to infinity. We estimated this constant and established that its value is $0.55080...$. The bounds were obtained on the computer, according to methods developed in the paper.

### *6.1.2. Efficient computation of local periods*

In continuation of our work on algorithmic methods for computing various repetition structures in words, this year we studied the problem of efficient computation of all local periods of a given word. For each position of the word, the local period is defined to be the minimal $k$ such that there exists a square of length $2k$ centered at this position. The notion of local period is very important in word combinatorics, as illustrated, e.g., by the fundamental Critical Factorization Theorem. An efficient computation of local periods remained an open problem. The main difficulty was that this computation didn't seem to follow from efficient and powerful algorithms for computing all maximal repetitions of the word[5].

In collaboration with R. Kolpakov (currently at the University of Liverpool) as well as J.-P. Duval, T. Lecroq et A. Lefebvre (all from the *Université de Rouen*) we were able to design a linear-time algorithm for computing *all* local periods in a word. The algorithm uses the technique for computing maximal repetitions[5] in a non-trivial way. The paper describing these results has been presented in 2003 to the international conference *Mathematical Foundations of Computer Science (MFCS)* [16].

We also point out that the journal version of the paper, joint with R. Kolpakov, on computing approximate repetitions in words appeared this year in *Theoretical Computer Science* [10].

### *6.1.3. Local alignment of DNA sequences*

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

The method of YASS is based on first 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 closely located (including overlapped) 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* [28]. 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 (of different type). Formed groups are then tested by trying to extend them to high-scoring local alignments.

Compared to the well-known BLAST tool [25] or to the more recent Pattern-Hunter algorithm [43], the YASS method for forming hits is more flexible and adapts to 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 can be achieved by using *spaced seeds* (see also 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). A complete description of YASS can be found in [23]. The work has been presented this year as a poster at RECOMB'2003 [17].

### 6.1.4. 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. In recent works [30][29], alignments were specified by a Markov model. The Markov model approach, however, allows to capture *local properties* of alignments but does not allow to specify their *global properties*.

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 effectively found by virtually all algorithms and that are expected to be found.

In collaboration with Yann Ponty (LRI Orsay), we have developed a probabilistic model for homogeneous sequences alignments, and proposed, on the one hand, 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. The conclusion of this study is that ignoring the property of homogeneity introduces a bias in measurement. This proves the importance of our approach. The paper describing this work [22] is submitted to a conference.

### 6.1.5. Approximate pattern matching using multiple seeds

Several algorithms and associated data structures for efficient pattern matching without errors are known. However, those methods do not extend to the case when a searched pattern is allowed to contain errors (e.g. letter substitutions).

Most of existing approximate pattern matching methods are based on the 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 that match exactly (or with one error) respective fragments of the pattern. More recently, methods based on spaced fragments have been introduced [31] and have been shown to improve the efficiency.

In this context, we have proposed a new method based on *multiple spaced seeds* that enables to increase the filter selectivity by a factor of more than 20 at the price of having additional hash tables for the text. The idea here is to combine multiple seeds to solve in a disjunctive way 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*.

An experimental program has been implemented for designing efficient multiple seeds. The main intended application of the program is the design of oligonucleotides for DNA chip experiments. Note that there is no need here to store several hash indexes in main memory. The program uses a genetic algorithm to search over seed families, in combination with fast dynamic programming methods for estimating the sensitivity. A paper describing this work is under preparation.

## 6.2. Discrete geometry

### *6.2.1. Noisy curves*

The recognition of digital objects is an important topic in discrete geometry and numerous works on discrete lines and arcs have been done (cf [34][49] ). We got interested in the notion of "fuzzy" (noisy) digital objects and in their detection. Note that 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. Fuzzy segments*

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

The definition of fuzzy segments relies on the definition of arithmetical discrete line given by J.P. Reveillès [46] in 1991: an arithmetical discrete line is a set of integer points $(x, y)$ verifying $\mu \leq ax - by < \mu + \omega$, where all parameters are integer values. The number $\frac{a}{b}$ represents the slope of the line, $\mu$ its position in the plane and $\omega$ its thickness.

A fuzzy segment is an 8-connected sequence of points that belongs to an arithmetical discrete line with a given thickness. A parameter, the order of a fuzzy segment, controls the level of the authorized noise via the thickness of the straight line bounding the fuzzy segment. Fuzzy segments generalize naive lines, where $\omega$ is fixed to $max(|a|, |b|)$, permitting at the same time that some points are missing.

Adding a point to a fuzzy segment is translated to the computation of the slope and of the thickness of the new bounding discrete line. We showed that this computation is easy to do. It leads to an incremental and very efficient algorithm for splitting a discrete curve into fuzzy segments of fixed order.

We wrote a paper on this subject and presented it in May 2003 at the IWCIA Conference in Palermo [15]. Since then, we submitted an extended version to a special issue of Discrete Applied Mathematics, this version is published as INRIA report [20].

*6.2.1.2. Estimation of tangents to a noisy discrete curve*

We propose a new notion of discrete tangent, adapted to noisy curves. It relies on the definition of discrete tangents given by Anne Vialard in 1996 [48], on the definition of fuzzy segment and on the algorithm of fuzzy segments recognition mentioned in the paragraph above.

We proposed a variant of the existing algorithm for fuzzy segment recognition. The new algorithm recognizes, in linear time, a fuzzy 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 fuzzy segment of order $d$ recognized by the algorithm at the point $P$. We then obtained an algorithm that computes the parameters of the approximate discrete tangent at each point of a discrete curve. From this algorithm, we can deduce several approximate parameters as the normal vector or the curvature at each considered point.

This work will be presented to the next Vision Geometry XII Conference to be held in January 2004 in San Jose as a part of the 16th IS&T/SPIE Annual Symposium.

### *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 a study of particular figures called hv-convex polyominoes. We developed a linear-time incremental algorithm to detect the convexity of such polyominoes. An extended version of the paper on this subject [3] was published in *Discrete Applied Mathematics* in January 2003 [8].

More recently, contacts were established with the University of Hamburg, namely with Professor U. Eckart and his student H. Reiter. In collaboration with the latter, 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 proved in [8] for the convex case, and actually extends them to the concave case. The paper describing this work is in preparation.

### 6.2.3. Digital plane recognition

A naive digital plane with integer coefficients is defined as the 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.

The program code is short and elementary (less than 300 lines) and available on the Web[3].

Although the theoretical complexity of the algorithm is bounded by $O(n^7)$, it is very efficient in practice. A paper describing this work [21] is submitted to Discrete Applied Mathematics.

# 8. Other Grants and Activities

## 8.1. 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*.

## 8.2. National Initiatives

Members of the team participate in the *Action Spécifiques Algorithmes et séquences (AS 77)*[4] and *Indexation de texte et découverte de motifs (AS 185)*[5] of the *CNRS*, in the framework of the *Réseau Thématique Pluridisciplinaire Bioinformatique (RTP 41)*. Gregory Kucherov is a co-animator of both these working groups. The first meeting of the group *Algorithmes et séquences* was held at Loria in Nancy on January 20-21, 2003.

Members of the team also participate in the *Action Spécifique Géométrie discrète et géométrie algorithmique*, the first meeting of the group will be held in January 2004.

We participated in a project proposition in response to the French national call IMPBio (*Informatique, Mathématiques et Physique pour la Biologie*), together with LIRMM, *Centre d'Ecologie Fonctionnelle et Evolutive* and *Institut de Génétique Humaine* of Montpellier. Although the project has not been accepted, it has been selected to the complementary list, with the encouragement of resubmission.

## 8.3. International Initiatives

Our team participates in an *Arc-en-Ciel* collaborative project with the University of Haifa in Israel on the subject of DNA curvature.

We have numerous "informal" international collaborations, as illustrated by visits of foreign researchers to our group (Section 8.4) as well as by visits of members of the group to international institutions (Section 9.3).

## 8.4. External visitors

Alexandre Bolshoy, professor of the University of Haifa, visited our group for one week in September 2003, in the framework of our common *Arc-en-Ciel* project.

[3] http://www.loria.fr/~debled/plane/
[4] http://www.loria.fr/~noe/AlgSeq/algseq.html
[5] http://degas.lirmm.fr/ASIM/

Helene Reiter-Doerksen, a PhD student from the University of Hamburg (Germany) made a three-weeks visit to our team, within the Sokrates-Erasmus program. During the stay, she worked on polygonal decomposition of discrete sets.

Mikhail Roytberg, senior researcher of the Institute of Mathematical Problems in Biology in Puschino (Russia), visited our team for one month in the fall 2003 as an INRIA invited professor. He gave a seminar *From an Analysis of Protein Structural Alignments Towards a Novel Approach to Align Protein Sequences* at the LORIA bioinformatics seminar.

# 9. Dissemination

## 9.1. Services

G. Kucherov served on the program committee of the 3rd Workshop on Algorithms in Bioinformatics (Budapest, Hungary, September 2003) and the Andrei Ershov 5th International Conference Perspectives of System Informatics (Novosibirsk, Russia, July 2003).

I. Debled-Rennesson is a full member of the *commission de spécialistes scientifique de l'IUFM de Lorraine* and a deputy member of the *commission de spécialistes en 27ième section de l'université de Reims Champagne-Ardenne*. She is a member of technical committee IAPR on discrete geometry (TC18). In October 2003, she was elected at the *Conseil National des Universités (CNU), 27ième section*.

J. Rouyer is a vice-president of the *commission de spécialistes en 27e section* of the *Université Henri Poincaré de Nancy*.

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

## 9.2. Teaching

I. Debled-Rennesson and G. Kucherov supervised the internship of Patricia Lavigne (DESS *EGOIST* of Rouen) from November 2001 to July 2003.

I. Debled-Rennesson et L. Noé supervised the internship of Ougas Elmi Houssein (DESS *Compétences Complémentaires en Informatique*) in June-September.

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 to the DESS *Ressources Génomiques et Traitements Informatiques* of the *Université Henri Poincaré de Nancy*.

In the framework of their teaching workload, L. Noé, J. Rouyer and I. Debled-Rennesson delivered various computer science courses at the *Université Henri Poincaré de Nancy* (DESS CCI, ESIAL, MIAS2), as well as at the *IUFM de Lorraine*. J. Rouyer is also responsible for the specialization *Ingénierie du Logiciel* of *ESIAL*.

J. Rouyer and I. Debled-Rennesson supervised three ESIAL students in a research project on discrete geometry: computing possible centers for circles separating two sets of points.

## 9.3. Participation in meetings, seminars, invited talks

### *9.3.1. Meetings, tutorials, conferences, invited seminar talks*

L. Noé participated in the School of young researchers *Ecole Jeunes Chercheurs en Algorithmique et Calcul Formel* that was held in Paris in March 2003. He gave a talk there on heuristic methods of local alignment. He also participated at the *International Summer School in Computational Biology* that was held in Warsaw in September 2003.

G. Kucherov, H. Zganic and L. Noé organized a meeting of the working group *Algorithmes et Séquences* at LORIA on January 20-21, 2003. G. Kucherov and L. Noé both gave a talk at this meeting on **mreps** 2.5 and YASS software respectively.

G. Kucherov, I. Debled-Rennesson, and L.Noé participated in the RECOMB03 conference in Berlin in April 2003. G. Kucherov and L.Noé presented a poster at that conference.

G. Kucherov, I. Debled-Rennesson, F.Touzain and L.Noé participated in the international conference ECCB/JOBIM in Paris in October 2003. G. Kucherov, I. Debled-Rennesson, and F.Touzain presented a poster there.

L. Noé is invited to make a talk in December 2003 at the seminar *Mathématiques pour le Génome* in Evry.

G. Kucherov and L. Noé participated in the meeting of the *Action Spécifique Indexation de texte et découverte de motifs* in Montpellier held on 20-21 November 2003. They both made a talk there on respectively seed-based local alignment and filtering techniques for approximate string matching.

Besides what is mentioned above, G. Kucherov made the following talks during the last year:

- at the Max-Plank Institute for Molecular Genetics in Berlin in December 2002,
- at the bioinformatics seminar of the Warsaw University in December 2002,
- at LIAFA (Paris 7) in March 2003,
- at LRI (Paris 11) in April 2003,
- at the Moscow Conference on Computational Molecular Biology (MCCMB'03) in July 2003,
- at the Computational Biology Lab of the Brigham and Women's Hospital (Harvard Medical School) in Boston in August 2003,
- at the King's College London in October 2003.

P. Lavigne presented a poster at the *3ème Journées Post-Génomique de la DOUA (JPGD'03)*, held in May 2003.

J. Rouyer and I. Debled-Rennesson presented their work at the *International Workshop on Combinatorial Image Analysis (IWCIA 2003)* that took place in Palermo (Italy) in May 2003.

I. Debled-Rennesson made an invited talk at the French seminar on Discrete Geometry in Lyon in June 2003. In November, she also participated in the international conference *Discrete Geometry for Computer Imagery (DGCI)* in Naples.

### *9.3.2. Visits of team members*

G. Kucherov visited the Mount Sinai Medical School in order to work with Professor Gary Benson on integrating **mreps** software into the Tandem Repeat DataBase (TRDB) created in his group.

During August 2003, G. Kucherov visited the DIMACS center at Rutgers University in New Jersey (USA), invited by Professor Martin Farach-Colton.

## 9.4. Participation in juries

I. Debled-Rennesson participates in the jury of PhD thesis of X. Hilaire (LORIA) to be held in January 2004.

G. Kucherov participated in the jury of the DESS diploma of Patricia Lavigne. The defence was held in Rouen in June 2003.

# 10. Bibliography

## Major publications by the team in recent years

[1] I. DEBLED-RENNESSON. *Étude et reconnaissance de droites et plans discrets.* Thèse de doctorat, Université Louis Pasteur, Strasbourg, Décembre, 1995.

[2] I. DEBLED-RENNESSON, J.-P. REVEILLÈS. *Incremental algorithm for recognizing pieces of digital planes.* in « Spie's International Symposium on Optical Science, Engineering, and Instrumentation, Technical Conference, Vision Geometry 5, Denver, USA », 1996.

[3] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Detection of the Discrete Convexity of Polyominoes.* in « DGCI'2000, Uppsala, Suède », series Lecture Notes in Computer Science, volume 1953, Springer-Verlag, pages 491-504, December, 2000.

[4] R. KOLPAKOV, G. KUCHEROV. *Finding Approximate Repetitions under Hamming Distance.* in « 9-th European Symposium on Algorithms (ESA 2001), Aarhus, Denmark », series Lecture Notes in Computer Science, volume 2161, F. AUF DER HEIDE, editor, pages 170 – 181, August, 2001.

[5] 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 (USA) », IEEE Computer Society, pages 596–604, New-York, 17-19 octobre, 1999.

[6] R. KOLPAKOV, G. KUCHEROV, Y. TARANNIKOV. *On repetition-free binary words of minimal density.* in « Theoretical Computer Science », number 1, volume 218, 1999.

[7] I. DEBLED-RENNESSON, J.-P. REVEILLÈS. *A linear algorithm for segmentation of digital curves.* in « International Journal of Pattern Recognition and Artificial Intelligence », 1995.

## Articles in referred journals and book chapters

[8] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Detection of the Discrete Convexity of Polyominoes.* in « Discrete Applied Mathematics », number 1, volume 125, Jan, 2003, pages 115-133.

[9] R. KOLPAKOV, G. BANA, G. KUCHEROV. *mreps : efficient and flexible detection of tandem repeats in DNA.* in « Nucleic Acids Research », number 13, volume 31, Jul, 2003, pages 3672-3678.

[10] R. KOLPAKOV, G. KUCHEROV. *Finding approximate repetitions under Hamming distance.* in « Theoretical Computer Science », number 1, volume 303, Jun, 2003, pages 135-156.

[11] G. KUCHEROV, P. OCHEM, M. RAO. *How many square occurrences must a binary sequence contain?.* in « The Electronic Journal of Combinatorics », number 1, volume 10, Jan, 2003.

[12] D. RODIONOV, A. VITRESHCHAK, A. MIRONOV, M. GELFAND. *Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes.* in « Journal of Biological Chemistry », number 42, volume 272, Oct, 2003, pages 41148-41159.

[13] D. RODIONOV, A. VITRESHCHAK, A. MIRONOV, M. GELFAND. *Regulation of lysine biosynthesis and transport genes in bacteria : yet another RNA riboswitch?.* in « Nucleic Acids Research », number 23, volume 31, Dec, 2003, pages 1-10.

[14] A. VITRESHCHAK, D. RODIONOV, A. MIRONOV, M. GELFAND. *Regulation of the vitamin B(12) metabolism and transport in bacteria by a conserved RNA structural element..* in « RNA », number 9, volume 9, Sep, 2003, pages 1084-1097.

## Publications in Conferences and Workshops

[15] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Segmentation of Discrete Curves into Fuzzy Seg-*

*ments.* in « Proceedings of the 9th International Workshop on Combinatorial Image Analysis (IWCIA'2003), Palermo, Italy », series Electronic Notes in Discrete Mathematics, volume 12, May, 2003.

[16] J.-P. DUVAL, R. KOLPAKOV, G. KUCHEROV, T. LECROQ, A. LEFEBVRE. *Linear-Time Computation of Local Periods.* in « 28nd International Symposium on Mathematical Foundations of Computer Science - MFCS'03, Bratislava, Slovakia », series Lecture Notes in Computer Science, volume 2747, Springer Verlag, B. ROVAN, P. VOJTAS, editors, pages 388-397, Aug, 2003.

[17] G. KUCHEROV, L. NOÉ. *YASS : similarity search in DNA sequences.* in « The Seventh Annual International Conference on Research in Computational Molecular Biology - RECOMB'03, Berlin, Germany », Apr, 2003, poster.

[18] P. LAVIGNE, I. DEBLED-RENNESSON, B. AIGLE, P. LEBLOND, G. KUCHEROV. *Identification of transcription factor binding sites in the genome of Streptomyces coelicolor A3(2).* in « Journées Post Génomique de la Doua 2003 (JPGD2003), Lyon, France », May, 2003, Poster.

[19] 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 », Sep, 2003, Poster.

## Internal Reports

[20] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Segmentation of Discrete Curves into Fuzzy Segments.* Rapport de recherche, number RR-4989, INRIA, Nov, 2003, http://www.inria.fr/rrrt/rr-4989.html.

[21] Y. GÉRARD, I. DEBLED-RENNESSON, P. ZIMMERMANN. *A fast and elementary algorithm for digital plane recognition.* Rapport de recherche, Nov, 2003.

[22] G. KUCHEROV, L. NOÉ, Y. PONTY. *Estimating seed sensitivity on homogeneous alignments.* Rapport de recherche, number 5047, Nov, 2003, http://www.inria.fr/rrrt/rr-5047.html, article soumis.

[23] L. NOÉ, G. KUCHEROV. *YASS: Similarity search in DNA sequences.* Rapport de recherche, number RR-4852, INRIA, Jun, 2003, http://www.inria.fr/rrrt/rr-4852.html.

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

## Bibliography in notes

[25] 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 », number 17, volume 25, 1997, pages 3389–3402.

[26] 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 », AAAI Press, pages 28-36, Menlo Park, California, 1994.

[27] T. BAILEY, M. GRIBSKOV. *Combining evidence using p-values: application to sequence homology searches.* in « Bioinformatics », volume 14, 1998, pages 48-54.

[28] G. BENSON. *Tandem repeats finder: a program to analyse DNA sequences.* in « Nucleic Acids Research », number 2, volume 27, 1999, pages 573–580.

[29] 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) », series Lecture Notes in Computer Science, volume 2812, Springer, R. P. G. BENSON, editor, September, 2003.

[30] 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, pages 67-75, April, 2003.

[31] S. BURKHARDT, J. KÄRKKÄINEN. *Better filtering with gapped q-grams.* in « Fundamenta Informaticae », number 1-2, volume 56, 2003, pages 51-70.

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

[33] C. CHELALA, M.-D. DEVIGNES, S. IMBEAUD, R. ZOOROB, C. AUFFRAY, E. CURIS, S. BÉNAZETH, D. COX. *Inconsistencies between maps of human chromosome 22 correlate with increased frequency of disease-related loci.* in « Journal of Biological Systems », number 4, volume 10, 2002, pages 303-317.

[34] D. COEURJOLLY, L. TOUGNE, Y. GÉRARD, J.-P. REVEILLÈS. *An Elementary Algorithm for Digital Arc Segmentation.* volume 46, 2001.

[35] M. CROCHEMORE. *Recherche linéaire d'un carré dans un mot.* in « Comptes Rendus Acad. Sci. Paris Sér. I Math. », volume 296, 1983, pages 781–784.

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

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

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

[39] A. FRAENKEL, J. SIMPSON. *How many squares must a binary sequence contain?.* in « Electronic Journal of Combinatorics », number R2, volume 2, 1995, pages 9pp, http://www.combinatorics.org/Volume_2/volume2.html#R2.

[40] F. GRUNDY, S. LEHMAN, T. HENKIN. *The L-box regulon: Lysine sensing by leader RNAs of bacterial lysine biosynthesis genes.* in « Proc. Natl. Acad. Sci. USA. », number 21, volume 102, 2003, pages 12057-62.

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

[42] D. KNUTH, J. MORRIS, V. PRATT. *Fast pattern matching in strings.* in « SIAM J. Comput. », volume 6, 1977, pages 323–350.

[43] B. MA, J. TROMP, M. LI. *PatternHunter: Faster and more sensitive homology search.* in « Bioinformatics », number 3, volume 18, 2002, pages 440-445.

[44] G. MANACHER. *A new linear-time on-line algorithm for finding the smallest initial palindrome of the string.* in « J. ACM », volume 22, 1975, pages 346–351.

[45] F. MURI-MAJOUBE, B. PRUM. *Une approche statistique de l'analyse des génomes.* in « Gazette des mathématiciens », volume 89, 2001.

[46] J.-P. REVEILLÈS. *Géométrie discrète, calculs en nombre entiers et algorithmique.* Thèse d'état, Université Louis Pasteur, Strasbourg, 1991.

[47] E. SHPIGELMAN, E. TRIFONOV, A. BOLSHOY. *CURVATURE: software for the analysis of curved DNA.* in « Comput Appl Biosci », number 4, volume 9, 1993, pages 435-40.

[48] A. VIALARD. *Geometrical Parameters Extraction from Discrete Paths.* in « 6th DGCI », series Lecture Notes in Computer Science, volume 1176, Springer-Verlag, pages 24-35, 1996.

[49] W. WAN, J. A. VENTURA. *Segmentation of Planar Curves into Straight-Line Segments and Elliptical Arcs.* in « Graphical Models and Image Processing », volume 59, 1997, pages 484–494.