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 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, DR CNRS from October 2005]

Administrative assistant

Céline Simon [TR INRIA]

Research scientists

Isabelle Debled-Rennesson [Maître de conférences, IUFM de Lorraine, *détachée* CR INRIA until September 2005]
Bertrand Kerautret [Maître de conférences, IUT Saint Dié, from September 2005]
Jean-Luc Rémy [CR CNRS, part time]

PhD students

Laurent Noé [grant MJENR, until September 2005]
Fabrice Touzain [grant INRIA co-sponsored by the Lorraine region]
Laurent Provot [grant MJENR, from September 2005]

Post-doctoral fellow

Alexey Vitreschak [INRIA, until October 2005]

Visiting scientists

Miklós Csűrös [Université de Montréal, June 1-30, 2005]
Anna Gambin [Warsaw University, August 22 - September 16, 2005]
Slawomir Lasota [Warsaw University, August 22 - September 16, 2005]
Mikhail Roytberg [Institut for Mathematical Problems in Biology, Russia, September 12 - October 31, 2005]
Ivan Tsitovich [Institute of Information Transmission, Russia, October 17-30, 2005]

Technical staff

Christophe Valmir [contractor engineer, January-October 2005]

Internships

Thanh-Phuong Nguyen [IFI, March-September 2005]
Mathilde Bouvel [ENS de Cachan, March-July 2005]
Sylvain Blondeau [July-August 2005]
Charlotte Lieunard [July-August 2005]

2. Overall Objectives

2.1. 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 pay 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 have been 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. Finally, several software programs (SIGFFRID, REPCLUSTER) are currently under development (see the software section below).

The year 2005 has been of a particular importance for ADAGE. On the one hand, it is the last year of the existence of the team. Due to the leave of two of its members, including the team leader, ADAGE will give rise to a new team in 2006. On the other hand, ADAGE was evaluated on March 31 - April 1, 2005, at the INRIA evaluation seminar for the BIO program. The evaluation of our work during 2001-2004 proved to be positive in the whole. The obtained results, both their originality and significance, have been appreciated by the experts. Even more important, the whole methodological approach of Adage has been approved. We therefore consider this evaluation to be a successful conclusion of the work of ADAGE.

3. Scientific Foundations

3.1. Scientific Foundations

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

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 [30], [35], [28], [29]. 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 [26]. 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.

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 4.1.4, 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.

In this work, we apply a comparative genomics approach, based on comparison of several phylogenetically related genomes. We developed a software, called SIGFFRID, to perform a comparative analysis of putatively orthologous genes between *two* genomes. The idea of the 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. Previously, we grouped interesting motifs by

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

similarities of pairs of trinucleotides. Recently, SIGFFRID were modified so as to allow groupings by pairs of oligonucleotides with fixed gaps. At the next step, we extend each found motif within the corresponding sequences coming from the same genome. This was done previously by a recursive procedure that analyses the alignment of all sequences which share a motif, and clusters sequences according to the proportion of each letter found at a given position in the alignment. We have improved it by replacing for extension criteria the proportions of a letter by the probability to obtain at less as many occurrences of this letter as we observe in the number of sequences used at this position. 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. We have performed this last test by verifying the significance of this ratio. It consists of a test of likelyhood ratio providing a measure of the overconcentration of our motif in intergenic upstream sequences comparatively to the whole genome (work in collaboration with Sophie Schbath, INRA Jouy-en-Josas). The version of this algorithm using probabilities and only pairs of trinucleotides for grouping was presented in a long talk at JOBIM 2005 [17].

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. Multy-copy repeats in genomic sequences

We developed a REPCLUSTER program for computing conserved elements in genomic sequences. In general, a DNA conserved element is a sequence element that appears at least in two approximate copies in the input sequence. There are several programs specifically devoted to the computation of pairwise repeats within a given genomic sequence [38], [44], [39]. On the other hand, such pairwise repeats can be obtained by computing local similarities between the input sequence and itself using a local alignment method. In order to identify multy-copy conserved elements, we developed a novel algorithm for clusterisation of pairwise repeats, realized in the REPCLUSTER program. A set of local alignments found by YASS (see Section 5.3), or some other sequence alignment program, is submitted to the input of the REPCLUSTER program. All sequences found by YASS are then grouped into clusters, with the goal that each cluster corresponds to the same genomic element (such as highly repeated sequences, mobile elements, non-coding genes, regulatory elements, etc). A method of *cores* is used for clusterisation. Its main idea consists in using most conserved parts of repeats, called cores, for controlling the clusterisation process.

The clusterization of sequences detected by YASS is made in two steps. The first step (pre-clustering) consists in processing all local alignments. This pre-clustering step groups together sequences that are strongly related: this is achieved by a search for sub-graphs (almost perfect cliques) in the graph in which nodes are sequences and edges are similarities. These initial clusters are starting points for further clusterisation and are essential for the stability of cores of clusters.

The second step is a neighbour-joining clusterization algorithm based on sequence cores. Using cores allowed us to avoid incorrect clusterisation (joining together non-related clusters) and to accurately detect repeated units. First, a graph is constructed with nodes corresponding to the initial clusters. An edge connects two nodes when at least one sequence from one initial cluster overlaps at least one sequence from another initial cluster (which verify rules of length relation and overlapping of cluster cores). Using the cores, the clusterisation step is defined as the following traversal of the set of clusters: (a) after constructing the set of initial clusters, choose a start initial cluster (the largest one), (b) iteratively join the current cluster with other connected clusters, (c) recompute cores and go to step (b) again. After the whole clusterisation and filtering are

completed, all clusters are sorted by the number of elements. Moreover, quick gapless alignment of sequences belonging to the same cluster is computed.

We run REPCLUSTER on bacteria genomes and obtained a number of clustered repeats, some of them with a known biological function (clusters corresponded to mobile IS-elements, tRNAs, rRNAs, known extended regulatory elements, etc). In a number of genomes, short highly repeated sequences are detected. For example, we found a cluster of sequences of about 26bp long, which is highly distributed (several hundreds copies) in pathogenic *Neisseria spp.*. This is a palindromic element with consensus CGTCATTCCCRCRnARgYGGGAATC. Several copies of this element can form a complex palindromic structure with a few hierarchical levels of alternative palindroms. The complex repeated element, revealed by our procedure, is located in non-coding regulatory regions often adjacent to genes involved in bacterial pathogenesis. This element could be involved in gene regulation or in genome rearrangements. Moreover, in some others pathogenic bacteria, we also found highly-repeated elements often to be located closely to genes involved in bacterial pathogenesis.

4.1.4. 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 been done during this year.

4.1.4.1. 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 [22]. There are several software programs for modelling the DNA curvature [43], [33], 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), the first such program for computing the DNA curvature has been developed in our group in 2004 as a result of a DEA work [42]. This program does not require any user-defined parameter and enables the detection of 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. This year we continued this work by integrating our last proposed improvements related to the computation of the discrete curvature of 3D noisy curves. A paper describing this work is in preparation. Moreover, we plan to develop a user-friendly interface that would allow biologists to use this software.

4.1.4.2. Computer prediction of H-NS regulon in Escherichia coli

Nucleotide-associated proteins in enterobacteria are required for gene regulation and organization of chromosomal DNA [34]. H-NS was described as a transcription factor and was shown to play a role in modifying the structure of chromosomal DNA [21]. H-NS protein is known to bind non-specifically to intrinsically curved DNA and to have several cooperating binding sites. Moreover, H-NS binding regions contain several extended highly AT-rich segments. H-NS is involved in a negative control of expression of many unrelated bacterial genes. Many genes require activators to overcome the H-NS repression (FIS, IHF, OmpR, ArcA, RovA, dsrA(RNA), Lrp, ToxR, etc.). Despite of this, the function of H-NS in bacterial metabolism remains unclear. First and unique attempt to describe the H-NS regulon was made by Hommains et al. [36]. They showed that nearly 5% of the E. coli 4290 CDSs was altered on DNA arrays in the hns mutant strain. To elucidate and define more exactly the H-NS regulon in E. coli (direct gene regulation by H-NS), we compared DNA array data with data based on the prediction of DNA curvature and highly AT-rich segments in gene upstream regions.

We applied the CURVATURE program [43] and found all strong curved sites with the curvature value more than 0.192 in upstream regions of all *E.coli* genes. This threshold curvature value corresponds to significant curvature (average curvature value = 0.099, sigma=0.05, window size of 100 bp). The presence of AT-rich segments (HATS) (100-300 nucleotides long and with A/T content about 80%) is a feature of H-NS binding

regions. For the prediction of highly AT-rich segments we developed an algorithm based on the maximum scoring method. On average, the AT-content of *E.coli* upstream regions is 57%. Extended AT-rich segments with AT-content about 80% occur relatively rarely. All genes previously found by DNA array can be grouped in to 108 known or possible operons. 34 of these operons have both HATS and curvature sites. These operons are the best candidates for the direct H-NS regulation and deserve a further experimental study.

4.1.5. Transposable elements in plant genomes

During the visit of A. Gambin and S. Lasota (see Section 7.4), we worked on the prediction and analysis of transposable elements in *Medicaco Trancatula* genome. The starting point of this research has been a new autonomous DNA transposon found in this genome, named MTMaster, that has a transposase sequence homologous to that of the known DCMaster element. Starting from this finding, we made an exhaustive computer analysis looking for other copies of this transposon and found 21 of those.

For all those sequences, we predicted coding regions using the FGENESH[2] program. Each of the sequences contains at least two exons coding for the transposase. It has been found that a third exons often appears in the region coding the transposase (in 9 among the 21 sequences). In addition, in 15 cases a coding region homologous to the first ORF has been found. Orientation and ordering of both ORFs seems to be variable from one sequence to another.

We also mined the *Medicaco Trancatula* genome for other members of MtMaster family, namely for non-autonomous elements (MITEs). Based on subterminal regions of autonomous transposable elements, we identified 77 sequences overall, predicted to belong to this family of transposons.

A further analysis has been performed on those sequences, including an analysis of their relative position to genes and a phylogenetic analysis, both on the protein and DNA level. All those results are now being prepared for publication, in collaboration with polish biologists.

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.

[2] http://www.softberry.com/

5.2. mreps

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

mreps [37] 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://bioinfo.lifl.fr/mreps/
- from the Web page of INRIA free software http://www.inria.fr/valorisation/logiciels/index.fr.html

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)*[3] 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.*

Since 2003, we develop YASS – a software for computing similarity regions in genomic sequences (local alignment). YASS is more sensitive than the commonly used BLAST program, 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 below).

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.

During this year, we carried on the development of YASS software. The main improvement was the introduction of multiple seeds, that leads to a further improvement of the speed/sensitivity ratio. For example, using two seeds of weight 10 results in both a more sensible and a more selective search than the previous version of YASS that used a single seed of weight 9. In practice, this results in a 20% speed-up of the execution time and more complete set of results.

Except for multiple seeds, a number of more technical modifications has been done:

- the computation of Λ and K values of the Gumbel law has been improved according to the algorithm by Altschul et Karlin (1990). These two values are key parameters in estimating the p-value and E-value of similarities found by YASS.
- the use of cash-memory has been optimized; new hash tables have been added to speed up the computation of seed groups.
- a parallelization of the program via multi-threads has been introduced; this enables a gain in efficiency on biprocessor architectures but also on dual-core or hyperthreaded processors.
- the computation of alignment scores has been improved using the IUPAC alphabet for DNA representation as well as adequate scoring matrices.

The web server of YASS has been substantially improved :

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

- the interface of choosing/downloading the sequences has been improved,
- multiple seeds option has been integrated,
- format of output pages have been improved,
- session management (in order to avoid possible collisions of sumaltaneous executions) has been introduced.

A description of YASS software and of its web server appeared this year in *Nucleic Acid Research* [11].

6. New Results

6.1. Word combinatorics and algorithms on sequences

6.1.1. Repetitions in words

The book *Applied combinatorics of words* of Lothaire series was published in 2005 by Cambridge University Press. It contains a chapter by G. Kucherov and R. Kolpakov that presents numerous algorithmic and combinatorial results on repetitions (periodicities) in words, obtained by the authors for the last several years [9].

6.1.2. Local alignment of DNA sequences

Comparing biological sequences in order to find similar regions is at the core of bioinformatics and the corresponding software, such as the well-known BLAST [20] package, is by far the most widely used bioinformatics software. For the last three years, we have been developping YASS – a software program for similarity search in DNA sequences (see Section 5.3). Compared to BLAST, YASS benefits from two fundamental improvements.

The first one is the use of *spaced seeds*, first used for DNA similarity search by the PATTERNHUNTER algorithm [40]. YASS goes further with this approach and introduces a generalized seed model, called *transition-constrained seeds*. These seeds are an instance of a more general concept, called *subset seeds*, for which we developed a theoretical foundation and corresponding algorithms (see next section).

The second improvement of YASS concerns the so-called *hit criterion*, i.e. the way that the seeds are used to detect potential similarity regions. Here, we proposed a so-called *group criterion* that defines groups of seeds which are most likely to represent a single similarity region, according to a statistical (Bernoulli) sequence model. This further improves the sensitivity of the search.

A description of both improvements appeared previously in [41]. A more technical description of YASS software and of its web server appeared this year in [11].

6.1.3. Estimation of seed sensitivity

About three years ago 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?

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 [24], [23] 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. This automaton and the whole associated algorithm has been implemented in the HEDERA software[4].

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 has been presented to the WABI conference this year [16] and is now accepted for publication in *Journal of Bioinformatics and Computational Biology*.

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. An year ago, we proposed a new filtration technique based on *multiple spaced seeds* that extends the method introduced in [25] and enables a considerable increase of the filter selectivity (see the annual report of 2004). An extended description of this work appeared this year in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* [10].

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 [27], [45]. 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 digital 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 of 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 published in a special issue of Discrete Applied Mathematics [4].

In collaboration with Fabien Feschet from LLAIC (Clermont-Ferrand), we also proposed new results on a restriction of the class of blurred segments in order to guarantee the optimality in the recognition process. This work was presented at the International Conference *Discrete Geometry for Computer Imagery (DGCI)* [13].

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 use fixed parameters and automatically provides a partitioning of a discrete curve into its meaningful parts.

This work was published this year [5] in the *Electronic Letter on Computer Vision and Image Analysis (ELCVIA)*.

[4] http://www.loria.fr/projets/YASS/hedera.html

6.2.1.3. Discrete curvature

We proposed a new notion of discrete tangent, relying on the definition of blurred segments and adapted to noisy curves. An algorithm permits to calculate the parameters of the tangent at each point of a discrete curve. From this algorithm, we can calculate several parameters as the normal vector or the curvature at all points of the considered curve [31]. These results have been improved in dimension 2 and extended to dimension 3 with the notion of 3D blurred segment in the framework of a DEA work [18].

In bioinformatics, we are interested in the computation of the DNA curvature. Discrete models of the representation of the 3D structure of the DNA require the development of specific discrete geometry algorithms. We adapted to this problem the discrete curvature algorithms. The development of a software program is in progress.

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

The study of 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 [32].

Several 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 [32], and extends them to the concave case. A paper describing this work was published by Kluwer in a book [6].

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). A short program code (less than 300 lines) solving the problem is available on the Web[5] and a paper describing this work is published to *Discrete Applied Mathematics* [7].

We are now interested in the recognition of noisy discrete planes. This year, this topic was a subject of a DEA work done in our group [19]. In this work, techniques of discrete geometry were combined with those of computational geometry. A new definition was introduced: blurred discrete pieces of a plane. We have shown that the problem of recognition of these objects is equivalent to the one of computation of the thickness of a set of points in dimension 3. A corresponding algorithm was proposed. This study was presented at the *Journées Informatique et Géométrie* in October and is continued in the PhD work of L. Provot. The general goal of this work is to study the algorithms of recognition of pieces of naive discrete planes and their adaptation to the pieces of blurred planes in order to be used in the polyhedrisation of noisy discrete objects.

6.2.4. Discrete surface reconstruction from shading images

We introduced in [3] a discrete approach to the reconstruction of discrete surfaces from shading images. The main idea of this approach is to combine geometric information of the discrete surface to be reconstructed with photometric information. When only one light source oriented in viewer direction is used for the reconstruction, the method allows to add explicit constraints in order to reduce the concave/convex ambiguity.

Since this discrete approach does not use an analytical expression of the reflectance map (usually Lambertian), the reconstruction was applied with other reflectance map such as the specular Nayar's model. Furthermore, this method presents the advantage not to be limited to the uniform light source model. It can be easily extended with a non-distant light source.

[5] http://www.loria.fr/~debled/plane/

6.2.5. Discrete surface smoothing

In collaboration with archeologists, we have been working on the problem of geometric smoothing and parameters extraction of discrete surfaces using the approach described in the previous section.

A statistical and geometrical method to smooth discrete surfaces was introduced in [14]. This method consists in smoothing the object surface by moving the center of each voxel to the unit cube according to the projection to the tangent plane. The tangent plane was estimated by a statistical estimation and by considering geometric constraints. The resulting surface representation allows us to get both smooth normal vectors of the surface and a smooth mesh while preserving the geometrical properties of the surface.

This work was recently accepted in an extended version in the special issue of DGCI in *Computer & Graphics* [8].

6.2.6. Shape Modeling from Shading Design

Shading has a large impact to the human perception of 3D objects. Thus, in order to create or to deform a 3D object, it seems natural to manipulate its perceived shading. We proposed a new solution to implement this idea. Our approach is based on the ability of the user to coarsely draw a shading, under different lighting directions. With this intuitive process, the user can create or edit a height field (locally or globally), that will correspond to the drawn shading values. This approach is described in [15] and we are going to work on the extension of this approach in order to edit and create full 3D objects.

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

We are a part of the project REPEVOL funded by the ACI IMPBio program (*Action Incitative Coopérative "Informatique, Mathématiques et Physique pour la Biologie"*) of the French government. This is a joint project with LIRMM, *Centre d'Ecologie Fonctionnelle et Evolutive* and *Institut de Génétique Humaine* of Montpellier, and Boston University, USA. We are also a part of the working group *Algorithmique Génomique* funded by the MATHSTIC program of CNRS.

7.3. International Initiatives

Two international collaboration projects have been accepted in 2005. Our collaboration with the bioinformatics group of the Warsaw University (J. Tiuryn, A. Gambin, S. Lasota) has been "formalized" through a *Polonium* project accepted in 2005 for two years. We also have an active collaboration with the Institute of Mathematical Problems in Biology in Puschino, Russia (group of M. Roytberg). A three-partite collaboration with these polish and russian groups resulted in an *ECO-NET* project, funded in 2005 and submitted for prolongation in 2006. A number of visits have been made in 2005 within these two projects (see the next section).

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

7.4. External visitors

Miklós Csűrös, researcher from the University of Montreal, visited our group for one month in June as an INRIA invited professor. During this visit, we were doing a collaborative work on probe design and annotation pf repeat regions of genomes.

Within the Polonium and ECO-NET projects (see previous section), the following visits took place in 2005:

- Ania Gambin and Slawomir Lasota, both lecturers at Warsaw University, stayed for one month with our group in August-September. During this stay, we undertook a joint work on the prediction of transposable elements in plans genomes (see section 4.1.5). Ania Gambin made a talk at the bioinformatics seminar of LORIA.
- Mikhail Roytberg, senior researcher of the Institute of Mathematical Problems in Biology in Puschino (Russia), visited our team for one month in October.
- Ivan Tsitovich, senior scientist of the Institute of Information Transmission in Moscow, came to visit us for two weeks in October.

The following researchers were invited by our group to make a seminar at the bioinformatics or algorithmics seminars of LORIA: Isabelle Sivignon (Laboratoire des Images et des Signaux, INPG), Maurice Margenstern (LITA, Metz), Joan Hérisson (LIMSI-CNRS), Vincent Bassano (LaMI, Université d'Evry), Yukiko Kenmochi (CNRS, Marne-la-Vallée). Martin Charles Golumbic, a professor from Haifa University, made a one-day visit to our group in June 2005.

8. Dissemination

8.1. Services

G. Kucherov served on the program committees of the 2nd Moscow Conference on Computational Molecular Biology (MCCMB'05, July 2005) and the 5th Workshop on Algorithms in Bioinformatics (Palma de Mallorca, Spain, October 2005). He now participates in the program committee of the 6th International Andrei Ershov Memorial Conference *Perspectives of System Informatics* (PSI'06) that will take place in Novosibirsk (Russia) July 2006.

I. Debled-Rennesson served on the program committees of the 12th International conference on Discrete Geometry in Computer Imagery (Poitiers, France, April 2005). She is a member of the IAPR technical committee on discrete geometry (TC18) [6].

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*.

G. Kucherov has been, until October 2005, a member of the *Commission de Spécialistes* of the *Université Henri Poincaré Nancy 1*.

J.-L. Rémy is a member of the *Conseil du Laboratoire* of LORIA. He is assigned to syndical activities within 30% of annual service.

8.2. Teaching

I. Debled-Rennesson supervised the DEA training of Laurent Provot in February-August 2005, and the DEPA (*Diplôme d'études professionnelles approfondies*) training of Phuong Nguyen Thanh (IFI, Hanoi) in March-September 2005.

G. Kucherov supervised the internship of Mathilde Bouvel (*Ecole Nationale Supérieure de Cachan*) in April-July 2005.

F. Touzain supervised summer internships of Charlotte Lieunard and Sylvain Blondeau, both students of the University of Nancy (*Maîtrise de Biologie Cellulaire et Physiologie*).

In the academic year 2004-2005, G. Kucherov taught the course *Algorithmics of discrete structures* of *DEA d'Informatique* of Nancy (jointly with D. Kratsch, *Université de Metz*). 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*.

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

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

L. Provot, B. Kerautret and I. Debled-Rennesson participated in the *Journée informatique et géométrie* in october at Paris. L. Provot gave a talk at this meeting.

F. Touzain and I. Debled-Rennesson participated in the JOBIM conference in Lyon in July. F. Touzain gave a talk at this conference.

I. Debled-Rennesson participated in a one-week school *Introduction avancée à la biologie post-génomique* in September in Evry. She also participated in the 12th International Conference DGCI in Poitiers in April, where she co-authored a presentation, jointly with a member of LLAIC (Clermont-Ferrand).

G. Kucherov, L. Noé, M. Roytberg and A. Vitreschak participated in the 2nd Moscow Conference on Computational Molecular Biology in July 2005. G. Kucherov, M. Roytberg and A. Vitreschak gave talks at that conference.

G. Kucherov and M. Roytberg participated in the 5th Workshop on Algorithms in Bioinformatics (Palma de Mallorca, October 2005). G. Kucherov gave a talk there.

G. Kucherov attended the 9th Annual International Conference on Research in Computational Molecular Biology (RECOMB) that took place in Boston in May 2005.

Besides, G. Kucherov gave the following talks:

- at *London Stringology Days* in February 2005,
- at the LIFL seminar in Lille in February 2005,
- at the 2nd Haifa Annual International Stringology Research Workshop of the Israeli Science Foundation in April 2005,
- an invited talk at the 31st International Workshop on Graph-Theoretic Concepts in Computer Science (WG) in Metz in June 2005 (an extended abstract of this talk is to appear in [12]),
- at the meeting of the working group *Algorithmique génomique* of the MathSTIC program of CNRS in Orsay in November 2005,
- an invited talk at the LIX Workshop 2005 *Bioinformatics: algorithms, structures and statistics* in Palaiseau in December 2005.

L. Noé attended the meeting of the working group *Algorithmique génomique* of the MathSTIC program of CNRS in Orsay in November 2005.

8.3.2. Visits of team members

G. Kucherov made a one-month stay in Moscow in July-August 2005, funded by the ECO-NET project. Besides participating in the MCCMB conference during that stay, he continued his collaboration with M. Roytberg as well as with other russian colleagues (R. Kolpakov, A. Mironov, I. Tsitovich).

8.4. Participation in juries

G. Kucherov participated in the jury of PhD theses of François Nicolas and Denis Bertrand, both at LIRMM, Montpellier, in December 2005. He is also a reviewer of the habilitation thesis of Mathieu Raffinot, planned to be defended in February 2006 at the University of Marne-la-Vallée.

I. Debled-Rennesson participated in the jury of PhD thesis of D. Jamet in December at LIRMM, Montpellier.

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, January 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, July 2003, p. 3672-3678.

Articles in refereed journals and book chapters

[3] A. BRAQUELAIRE, B. KERAUTRET. *Reconstruction of Lambertian Surfaces by Discrete Equal Height Contours and Regions Propagation*, in "Image and Vision Computing", vol. 23, n° 2, February 2005, p. 177-189.

[4] I. DEBLED-RENNESSON, J.-L. RÉMY, J. ROUYER-DEGLI. *Linear Segmentation of Discrete Curves into Fuzzy Segments*, in "Discrete Applied Mathematics", vol. 151, n° 1-3, October 2005, p. 122–137.

[5] I. DEBLED-RENNESSON, S. TABBONE, L. WENDLING. *Multiorder polygonal approximation of digital curves*, in "Electronic Letters on Computer Vision and Image Analysis", Special Issue on Document Analysis, vol. 5, n° 2, August 2005, p. 98–110.

[6] H. DÖRKSEN-REITER, I. DEBLED-RENNESSON. *Convex and Concave Parts of Digital Curves*, in "Geometric Properties from Incomplete Data", R. KLETTE, R. KOZERA, L. NOAKES, J. WEICKERT (editors). , Computational Imaging and Vision, vol. 31, Springer-Verlag, January 2005.

[7] Y. GÉRARD, I. DEBLED-RENNESSON, P. ZIMMERMANN. *An elementary digital plane recognition algorithm*, in "Discrete Applied Mathematics", vol. 151, n° 1-3, October 2005, p. 169–183.

[8] B. KERAUTRET, A. BRAQUELAIRE. *A Reversible and Statistical Method for Discrete Surfaces Smoothing*, in "Computer & Graphics", (in press), vol. 30, n° 1, 2006.

[9] R. KOLPAKOV, G. KUCHEROV. *Identification of periodic structures in words*, in "Applied combinatorics on words", J. BERSTEL, D. PERRIN (editors). , Lothaire books, vol. Encyclopedia of Mathematics and its Applications, vol. 104, chap. 8, Cambridge University Press, 2005, p. 430–477.

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

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

Publications in Conferences and Workshops

[12] M. BOUVEL, V. GREBINSKI, G. KUCHEROV. *Combinatorial search on graphs motivated by bioinformatics applicaitons: a brief survey*, in "Proceedings of the 31st International Workshop on Graph-Theoretic Concepts in Computer Science (WG), Metz (France), June 23-25, 2005", D. KRATSCH (editor). , Lecture Notes in Computer Science, vol. 3787, Springer Verlag, 2005, p. 16–27.

[13] I. DEBLED-RENNESSON, F. FESCHET, J. ROUYER-DEGLI. *Blurred Segments Decomposition in Linear Time*, in "Proceedings of the 12th International Conference on Discrete Geometry for Computer Imagery, Poitiers, France", E. ANDRES, G. DAMIAND, P. LIENHARDT (editors). , LNCS, vol. 3429, Springer-Verlag, April 2005, p. 371-382.

[14] B. KERAUTRET, A. BRAQUELAIRE. *A Statistical Approach for Geometric Smoothing of Discrete Surfaces*, in "Proceeding of the International Conference on Discrete Geometry for Computer Imagery, Poitiers, France", E. ANDRES, G. DAMIAND, P. LIENHARDT (editors). , LNCS, vol. 3429, Springer-Verlag, April 2005, p. 405-414.

[15] B. KERAUTRET, X. GRANIER, A. BRAQUELAIRE. *Intuitive Shape Modeling by Shading Design*, in "Proceedings of the 5th International Symposium on Smart Graphics, Frauenwoerth Cloister, Germany", LNCS, n° 3638, Springer-Verlag, August 2005, p. 163-174.

[16] G. KUCHEROV, L. NOÉ, M. ROYTBERG. *A Unifying Framework for Seed Sensitivity and Its Application to Subset Seeds*, in "Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI), Mallorca (Spain), October 3-6, 2005", R. CASADIO, G. MYERS (editors). , Lecture Notes in Computer Science, vol. 3692, Springer-Verlag, 2005, p. 251–263.

[17] F. TOUZAIN, S. SCHBATH, I. DEBLED-RENNESSON, B. AIGLE, P. LEBLOND, G. KUCHEROV. *SIGffRid : Programme de recherche des sites de fixation des facteurs de transcription par approche comparative*, in "JOBIM", 2005, p. 417-425.

Miscellaneous

[18] T. NGUYEN. *Optimisation du calcul de la courbure de l'ADN*, DPEA report, IFI, October 2005.

[19] L. PROVOT. *Reconnaissance de morceaux de plans discrets bruités*, DEA report, LORIA, June 2005.

Bibliography in notes

[20] 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.

[21] T. ATLUNG, H. INGMER. *H-NS: a modulator of environmentally regulated gene expression*, in "Mol Microbiol", vol. 24, n° 1, April 1997, p. 7-17.

[22] 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.

[23] 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)", G. BENSON, R. PAGE (editors). , Lecture Notes in Computer Science, vol. 2812, Springer, September 2003.

[24] 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.

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

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

[27] 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.

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

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

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

[31] 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.

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

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

[34] C. DORMAN, P. DEIGHAN. *Regulation of gene expression by histone-like proteins in bacteria*, in "Curr Opin Genet Dev", vol. 13, n° 2, April 2003, p. 179-84.

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

[36] F. HOMMAIS, E. KRIN, C. LAURENT-WINTER, O. SOUTOURINA, A. MALPERTUY, J.-P. LE CAER, A. DANCHIN, P. BERTIN. *Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic*

nucleoid-associated protein, H-NS, in "Mol Microbiol", vol. 40, n° 1, April 2001, p. 20-36.

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

[38] 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, November 15 2001, p. 4633-42.

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

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

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

[42] 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.

[43] 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.

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

[45] 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.