INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Algo

Algorithms

Rocquencourt

THEME SYM

Activity Report

2005

Table of contents

1. Team

Head of project-team

Bruno Salvy [DR]

Vice-head of project team

Philippe Flajolet [DR]

Administrative assistant

Virginie Collette [TR]

Research scientists (Inria)

Alin Bostan [CR]
Frédéric Chyzak [CR]
Mireille Régnier [DR]

Research scientists (partners)

Philippe Dumas [professeur, Cl. Prépa. lycée Jean-Baptiste Say]
Pierre Nicodème [CR CNRS, École polytechnique]
Brigitte Vallée [DR CNRS, University of Caen]

Visiting scientists

Micha Hofri [starting September]

Ph. D. students

Julien Fayolle
Éric Fusy
Frédéric Giroire
Carine Pivoteau [starting October]
Vincent Puyhaubert [until March]
Antonio Vera [starting October]

Students intern

Carine Pivoteau [University of Paris VI, from March to September]
Antonio Vera [École polytechnique, from March to July]

Technical Staff

André Derrick Balsa [since November]

2. Overall Objectives

2.1. Overall Objectives

The primal objective of the project, inherited from the former century, is the field of *analysis of algorithms*. By this is meant a precise quantification of complexity issues associated to the most fundamental algorithms and data structures of computer science. Departing from traditional approaches that, somewhat artificially, place emphasis on worst-case scenarii, the project focusses on average-case and probabilistic analyses, aiming as often as possible at realistic data models. As such, our research is inspired by the pioneering works of Knuth.

The need to analyse, dimension, and finely optimize algorithms requires an in-depth study of random discrete structures, like words, trees, graphs, and permutations, to name a few. Indeed, a vast majority of the most important algorithms in practice either "make bets" on the likely shape of input data or even base themselves of random choices. In this area we are developing a novel approach based on recent theories of combinatorial analysis together with the view that discrete models connect nicely with complex-analytic and asymptotic methods. The resulting theory has been called *"Analytic combinatorics"*. Applications of it have been or are currently being worked out in such diverse areas as communication protocols, multidimensional search, data structures for fast retrieval on external storage, data mining applications, the analysis of genomic sequences, and data compression, for instance.

The analytic-combinatorial approach to the basic processes of computer science is very systematic. It appeared early in the history of the project that its development would greatly benefit from the existence of symbolic manipulation systems and computer algebra. This connection has given rise to an original research programme that we are currently carrying out. Some of the directions pursued include automating the manipulation of combinatorial models (counting, generating function equations, random generation), the development of "automatic asymptotics", and the development of a unified view of the theory of special functions. In particular, the project has developed the Maple library ALGOLIB, that addresses several of these issues.

3. Scientific Foundations

3.1. Analysis of Algorithms

Keywords: *analysis of algorithms*, *analytic combinatorics*, *asymptotic enumeration*, *combinatorial analysis*, *hashing methods*, *index tree*, *limit laws*, *random discrete structures*.

While we know the laws of basic physics and while probabilists have been setting up a coherent theory of stochastic processes for about half a century, the "laws of combinatorics", in the sense of the laws governing random structured configurations of large sizes, are much less understood. Accordingly, our knowledge in the latter area is still very much fragmentary. Some of the difficulties arise from the large variety of models that tend to arise in real-life applications—the world of computer scientists and algorithmic designers is really an artificial world, much more "free" than its physical counterpart. Some of us have then engaged in the long haul project of trying to offer a unified perspective in this area. The approach of analytic combinatorics has evolved from there.

Analytic combinatorics leads to discovering randomness phenomena that are "universal" (a term actually borrowed from statistical physics) across seemingly different applications. For instance, it is found that similar laws govern the behaviour of prime factors in integers, of irreducible factors in polynomials, of cycles in permutations, and of components in mappings of a finite set. Once detected, such phenomena can then be exploited by specific algorithms that factor integers (a problem relevant to public-key cryptography), decompose polynomials (this is needed in computer algebra systems), reorganize tables in place (this is of obvious interest in the manipulation of various data sets), and use collisions to estimate the cardinality of massive data ensembles. The underlying technology bases itself on generating functions, which exactly describe discrete models, as well as an interpretation of these generating functions as analytic transformations of the complex plane. Singularities together with the associated perturbative theory then deliver a number of very precise estimates regarding important characteristics of random discrete structures. The process can be largely made formal and accessible to computer algebra (see below) and it may be adapted to the broad area of analysis of algorithms.

3.2. Computer Algebra

Keywords: *Gröbner bases*, *asymptotic scales*, *random generation*, *special functions*.

Computer algebra at large aims at making effective large portions of mathematics, paying due attention to complexity issues. For reasons mentioned above, our project specifically investigates the way mathematical objects originating in complex analysis can be dealt with in an algorithmic way by computer algebra systems. Our main contributions in this area concern the automation of asymptotic analysis and the handling of special functions. The mathematical foundations of our algorithms are deeply rooted in differential algebra (Hardy fields for asymptotic expansions and Ore algebras for special functions).

Over the years, in order to automate the average-case analysis of ever larger classes of algorithms, we have developed algorithms and implementations for the following problems: the specification of formally specified combinatorial structures; the corresponding problems of enumeration and random generation; the automatic construction of asymptotic scales which is necessary for extracting the singular behaviour of generating functions; the automatic computation of asymptotic expansions in such scales; the automatic computation of asymptotic expansions satisfied by coefficients of generating series. An *Encyclopedia of Combinatorial Structures*, available on the web, gathers roughly one thousand structures for which generating series, recurrences, and asymptotic behaviour have been determined automatically using our libraries.

An important principle of computer algebra is that it is often easier to operate with equations defining a mathematical object implicitly rather than trying to obtain a "closed-form" expression of it. The class of linear differential and difference equations is particularly important in view of the large variety of functions and sequences they capture. In this area, we have developed the highly successful GFUN package (jointly with P. Zimmermann, from the Spaces project) dealing with the univariate case. In the multivariate case, we have developed the underlying theory based on Gröbner bases in Ore algebra, and an implementation in the MGFUN package. The algorithmic advances of the past few years have made it possible to start the implementation of an *Encyclopedia of Special Functions*, providing various information concerning classical functions (of wide use throughout sciences), including Bessel functions, Airy functions, ... The corresponding information is all automatically generated.

3.3. Algorithms on Sequences

Keywords: *combinatorics on words, genome, pattern matching, sequences.*

The goal of our research on sequences is the design of new algorithms and the computation of their average-case complexity or the derivation of combinatorial results on words and their implementation in statistical software. Possible applications are data compression and genomic sequences. A new area arises in the context of genomic sequences, where biologically significant motifs are extracted. This subject combines algorithms searching for potential signals (the candidates), and computations of statistical significance. For each candidate, the choice criterion is its underrepresentation or overrepresentation. Due to the large number of potential candidates, the speed and the numerical precision of the computation are crucial.

From a methodological point of view, we exhibit several renewal processes, and the limiting law is usually a Gaussian law. Here, the tail distributions are necessary, as one needs to evaluate the overrepresentation, or the underrepresentation, of a motif. The combinatorial properties of words allow, for this class of problems, an effective computation of formulae valid in the central domain and in the tails. Asymptotic analysis yields an exact expression of the rate function, in the sense of large deviation theory. Simultaneously, we define for each problems some characteristic languages in order to bound the computational complexity in the Markovian case.

4. Software

4.1. Software

The Algolib library is a set of Maple routines that have been developed in the project for more than 10 years. Several parts of it have been incorporated in the standard library of Maple, but the most up-to-date version is always available for free from our web pages. (The diffusion list for these updates contains more than 200 subscribers). This library provides: tools for combinatorial structures (the `combstruct` package), this includes enumeration, random or exhaustive generation, generating functions for a large class of attribute grammars; tools for linear difference and differential equations (the `gfun` package), which have received a very positive review in *Computing Reviews* and have been incorporated in N. Sloane's `superseeker` at Bell Labs; tools for systems of multivariate linear operators (the `Mgfun` package), including Gröbner bases in Ore algebras, that also treat commutative polynomials and are now the standard way to solve polynomial systems in Maple (although the user does not notice it); `Mgfun` has also been chosen at Risc (Linz) as the basis for their package `Desing`.

We also provide access to our work to scientists who are not using Maple or any other computer algebra system in the form of automatically generated encyclopedia available on the web. The Encyclopedia of Combinatorial Structures thus contains more than 1000 combinatorial structures for which generating series, enumeration sequences, recurrences and asymptotic behaviour have been computed automatically. The Encyclopedia of Special Functions gathers around 40 special functions for which identities, power series, asymptotic expansions, graphs, ...have been generated automatically, starting from a linear differential equation and its initial conditions. The underlying algorithms and implementations are those of GFUN and MGFUN. All the production process being automated, the difficult and expensive step of checking each formula individually is suppressed. Available on the web (http://algo.inria.fr/esf/), this encyclopedia also plays the rôle of a showcase for part of the packages developed in our project.

5. New Results

5.1. Analysis of algorithms

Participants: André Derrick Balsa, Philippe Flajolet, Éric Fusy, Frédéric Giroire, Vincent Puyhaubert, Mireille Régnier, Bruno Salvy, Brigitte Vallée.

There have been in 2005 two main streams of activity. One concerns general purpose methods for quantifying precisely properties of random discrete structures and algorithms. The other consists of applications to various areas of theoretical and practical computer science.

First, the general theory of analytic combinatorics, which serves as a basis to a modern vision of the average-case and probabilistic analysis of algorithms, has made progress with about 100 pages being written during the period (mostly on the general theory of complex asymptotic methods and meromorphic asymptotics in relation to regular languages, finite automata, and transfer matrices). The net result is a synthesis report of some 700 pages authored by Flajolet and Sedgewick [2] that is freely available on the web. It develops basic complex asymptotic methods from first elements of combinatorial theory (like in [28]) and results in a precise quantification of a great many properties of random discrete structures. This will result in a book to be published near the end of the year 2006 by Cambridge University Press.

Next, a number of algorithms of varied theoretical and practical interest have been conceived and/or analysed.

Data compression algorithms. Suffix trees are largely used as a data structure for representing texts in the realm of data compression (like in the gzip utility) and computational biology. Julien Fayolle has worked out the average-case behaviour of various parameters of the suffix tree, like size or external path length, under a memoryless source, thereby proposing an alternative to earlier approaches by Jacquet (HIPERCOM Project) and Szpankowski (Purdue). Recently, he has been able to extend these results to the broader class of Markovian models, and in a collaboration with M. Ward (Purdue), he has been able to characterize the expected depth of insertion in suffix trees. More recently, he has shown his methods to be applicable to a new scheme for data compresson due to Crochemore *et al.*, which, surprisingly enough, is based on maintaining an “antidictionary” (what gets encoded is a set of missing words!), and for which many characteristics had hitherto remained inaccessible to analysis. This last project represents joint work with Hiroyoshi Morita (University of Electro-Communications, Tokyo), Takahiro Ota (Nagano Prefectural Institute of Technology), and Philippe Flajolet. Julien Fayolle has completed the manuscript of his PhD thesis in December 2005, the defence being planned for March 2006.

Compact encoding of graphs. Éric Fusy has presented at the international conference SODA 05 (Symposium on Discrete Algorithms) a quasi-bijection between binary trees and 3-connected planar maps, found with D. Poulalhon and Gilles Schaeffer [25]. This bijection yields an encoding of 3-connected planar graphs, which corresponds to the true information content of polygonal meshes, and achieves optimal compression rate in the worst-case. Thanks to this bijection, it also becomes possible to generate uniformly at random 3-connected planar graphs in linear time, a task that is crucial in view of sampling planar graphs. Moreover, Éric Fusy has developed a new algorithm to draw a triangulation efficiently on a regular grid, upon relying on so-called transversal structures and face counting operations. The investigation of the combinatorics of the transversal structures makes possible the analysis of the size of the grid, which turns out to be almost surely $\frac{11n}{27} \times \frac{11n}{27}$ for a triangulation with n vertices, whereas previous algorithms gave an embedding on a $\frac{n}{2} \times \frac{n}{2}$ grid. This work [23] has been presented at the international conference Graph Drawing 2005.

Random generation and simulation. With P. Duchon (LaBRI), G. Louchard (Brussels), and G. Schaeffer (LIX), P. Flajolet has developed a brand new approach to the fast generation of complex structured configurations. The framework is inspired by Boltzmann models of statistical physics. Earlier studies (published in 2004) have treated exponential Boltzmann samplers, corresponding to labelled structures, that is, combinatorial objects with distinguishable atoms. This year, Carine Pivoteau (a fifth year student from MPRI who did her field study at INRIA Rocquencourt), Éric Fusy, and Philippe Flajolet have launched an ambitious programme meant to attack the general problem of unlabelled structures—this requires generating *in an unbiased way* objects that are invariant under symmetries. In particular, Carine Pivoteau has written her master thesis on this topic [29], and a joint study by Flajolet, Fusy, and Pivoteau is in preparation. The resulting algorithms are eminently practical: they are often linear (or quasi-linear) as regards computation time and make it possible routinely to generate objects of sizes near 100,000, while sizes only of the order of a few hundred were previously known to be attainable. Applications to random testing in software engineering and to the simulation of genomic data in computational biology are contemplated.

Random generation of planar graphs. It is known that 3-connected planar graphs are at the core of a decomposition of planar graphs. From this observation and using the powerful framework of Boltzmann samplers, Éric Fusy has been able to derive a very efficient random sampler for random labelled planar. Its time complexity is *linear*, as soon as a small tolerance on size is allowed, while the exact-size sampler has quadratic complexity. The previously known algorithms had complexity of the order of $O(n^7)$, and their memory requirements were prohibitive. With Fusy's algorithm, it now becomes possible to generate planar graphs of size about 10^5, which definitely constitutes a record. This work has been presented at the international conference AofA'05 (Analysis of Algorithms) [24].

Quantitative data mining. A major new discovery of the period 2003-2004 has been the LogLog-Counting algorithm of M. Durand and P. Flajolet. It permits us to estimate the cardinality (understood as the number of distinct records) in a huge file using a single pass and only about 2 kilobytes of auxiliary memory for an accuracy of about 1%. Frédéric Giroire has developed and thoroughly analysed a new promising category of algorithms based on order statistics. The algorithms apply to the gathering of a large number of simultaneous statistics on large "texts", which may equally well be natural language corpuses or router traces in networking. They are validated by a thorough mathematical analysis that combines several techniques developed in the project (e.g., generating functions, Mellin transforms, saddle-point methods). Within the framework of a national action on Massive Data Sets (ACI-MD), a project named FLUX is being supported. Flajolet, Giroire, and a newly recruited research engineer, André Balsa, are actively engaged in the design of high-performance cardinality estimators, which aim at being the best on the market. The problems involve fine system-code optimizations (Balsa and Chyzak), the selection of equitable hashing algorithms (Balsa and Giroire), as well as extensive testing on actual router traces. These activities are developed in close cooperation with our partners in FLUX, the RAP project of INRIA Rocquencourt and the Networking Team at LIRMM in Montpellier.

Work has also been ongoing regarding the emerging classification of combinatorial processes that are relevant to analysis of algorithms.

Coalescence. Phenomena involving Gaussian laws amongst discrete structures are by now fairly well understood, either through the classical theory of stochastic processes or within the framework of analytic combinatorics. Work conducted within the group has revealed next the importance of coalescences and confluences that are conducive to Airy phenomena. In this context, the computation of Gröbner bases, which is a very important subroutine of many computer algebra algorithms, turns out to be largely of a combinatorial nature and thus amenable to a treatment by the techniques of analytic combinatorics. In [18], under some technical conditions that avoid pathological cases, the complexity of Faugère's F_5 algorithm for Gröbner bases of overdetermined systems of degree 2 has been precisely quantified: It is polynomial in the number of solutions and is related to a quantity (the degree of regularity) that depends only on the number of equations and the dimension of the space (the number of unknowns). When the number of equations is slightly larger than the dimension, the growth of the degree of regularity involves the smallest zero of a Hermite polynomial; when the number of equations grows linearly with the dimension, its growth involves the largest zero of the Airy function. Work is in progress for the general case (arbitrary degrees).

Divide-and-Conquer analyses. For a great many probabilistic models encountered in discrete mathematics, singularities provide extremely precise and valuable information. The article [9] written by P. Flajolet in collaboration with J. Fill and N. Kapur (Johns Hopkins University) studies some classical tree models (binary search trees, Catalan trees, union-find trees) but not so standard toll functions. Functions amenable to singularity analysis are shown to be closed under Hadamard product (i.e., the termwise product of series). A valuable consequence is the possibility of classifying the solution to several basic recurrences of the probabilistic divide-and-conquer type whose central role in the design of efficient algorithms is well recognized.

Urn models. A new avenue to urn models has been opened when P. Flajolet, jointly with J. Gabarro and H. Pekari (Barcelona), have shown for the first time the possibility of developing a purely analytic model of urn processes of the Pólya type [12]. Theoretically, this reveals a classification of certain urn models based on the notion of genus and it leads to significant large deviation estimates, as well as to stable laws or to models exactly solvable in terms of elliptic functions in particular cases. In his PhD thesis [3] defended in March 2005, V. Puyhaubert completes the classification of 2×2 balanced urn models, while discovering extensions of the framework to 3×3 balanced urns, provided they are of triangular type. Such urn models can additionally describe classical and generalized coupon collector problems, balanced data structures of the B-tree type, as well as a simple model of conflicts (Flajolet and Puyhaubert, in preparation).

5.2. Computer Algebra

Participants: Alin Bostan, Frédéric Chyzak, Philippe Flajolet, Bruno Salvy.

There are two main directions in our research on linear differential and difference equations. One consists in developing new or faster algorithms dealing with the solutions of these equations. Another one consists in constructing equations known a priori to have such solutions, so that these fast algorithms can be applied.

This year, a new algorithm has been developed by A. Bostan, B. Salvy and Thomas Cluzeau (University of Limoges) for finding polynomial solutions of linear differential equations [20]. This is a basic operation that lies at the heart of many computer algebra algorithms for (definite or indefinite) integration. Even for the simple question of the existence, there is no general algorithm with polynomial complexity. However, the polynomial solutions of high degree are very much constrained by the equation and this makes it possible to represent them by a compact data structure: a recurrence and initial conditions. We have shown how this data structure can be exploited in order to compute polynomial solutions in quasi-optimal complexity (up to logarithmic factors).

The existence of fast algorithms for solutions of linear differential equations and recurrences makes it important to detect whether a function or a sequence is solution of such an equation. It is often possible to answer negatively by considering the asymptotic behavior of these functions or more generally their behavior or that of their generating series in the neighborhood of singular points. In [10], P. Flajolet, B. Salvy and Stefan Gerhold (RISC, Linz, Austria) have shown how a small set of techniques from analytic combinatorics can be applied with success in a variety of cases.

For several years, F. Chyzak and P. Paule (RISC, University of Linz, Austria) have been collaborating on the writing of a chapter on computer algebra methods for special functions, in the framework of the project Digital Library of Mathematical Functions (DLMF) of the National Institute of Standards and Technology (NIST). This ambitious project aims at providing a new edition for the "Handbook of Mathematical Functions," an authoritative handbook since 1962 and one of the most cited works in the history of scientific publications. The chapter is mainly concerned with those algorithms that are at the heart of our GFUN and MGFUN packages. A draft was finalized last year after interaction with NIST. The project—including edition by NIST, external validation, further interaction of NIST with the authors—has proved to be more time-demanding than first expected by NIST, and it has sustained several years of delay. The resulting product is expected to be published next year or in early 2007. The book will be available both in printed version (roughly 1,000 pages) and under electronic format (a CD and a web site, see http://dlmf.nist.gov/).

A recent follow up to F. Chyzak and B. Salvy's work is the application of methods originally developped for special functions to deal with symmetric functions in algebraic combinatorics. Together with Marni Mishna (UBC, Vancouver), they have devised algorithms for the computation of scalar products between symmetric series. For instance, this leads to: the efficient enumeration of classes of graphs given by regularity constraints; new symmetric functions identities with a representation-theoretic interpretation have been found; and new asymptotic results on the enumeration of regular graphs [7].

Another follow up of methods for special functions is an application to linear control systems. A collaboration of F. Chyzak with A. Quadrat (CAFE Project, INRIA Sophia-Antipolis) and D. Robertz (University of Aachen, Germany) has shown that elimination methods for non-commutative polynomials designed in the project permit to make methods developped by A. Quadrat for the recognition of properties of linear control systems effective. The spectrum of applications includes ODEs, PDEs, multidimensional discrete systems, differential time-delay systems, repetitive systems, multidimensional convolutional codes, etc. A package, OreModule, http://wwwb.math.rwth-aachen.de/OreModules/, has been developped, based on F. Chyzak's Maple implementation of Gröbner tools. The extended article [8] published this year develops the results sketched in two conference publications over the last two years.

For several years, B. Salvy and A. Bostan have been working jointly with the STIX laboratory of the *École polytechnique*. This work applies recent algorithmic progress on straight-line programs and has produced efficient algorithms and implementations for geometrical problems. Recently, this work has taken a new direction by extending to the numerical universe methods originally designed to deal with multiplicities when searching for symbolic solutions of polynomial systems. The results obtained by B. Salvy, G. Lecerf (University of Versailles Saint-Quentin-en-Yvelines), M. Giusti (École polytechnique) and J.-C. Yakoubsohn (University of Toulouse) are new versions of Newton's algorithm that are quadratically convergent even in the neighborhood of a multiple root or a cluster of roots for analytic functions [14] and polynomial systems [13] under a technical condition of "embedding dimension 1".

Now, the aim is to extend these methods based on geometric resolution to the non-commutative context necessary for the application to special functions. As a first step, it is necessary to obtain low complexity algorithms for algorithms based on evaluation and interpolation in the commutative case. Thus, in [6], A. Bostan and É. Schost (École polytechnique) have given sharp complexity estimates for the problems of multipoint evaluation and interpolation with respect to various polynomial bases and for special families of evaluation points, such as geometric and arithmetic progressions.

Another important operation on univariate polynomials is the recovery of the coefficients from the data of the first power sums of the roots. This is a crucial step for the algorithms by A. Bostan, P. Flajolet, B. Salvy and É. Schost in [5], which is devoted to the fast computation of some special bivariate resultants occurring in manipulations with algebraic numbers. Over fields of large characteristic, the use of fast exponentiation of power series is classical to perform the conversion from power sums to coefficients. This technique does not apply in small characteristic, because of divisions; A. Bostan, L. González-Vega (University of Santander, Spain), H. Perdry (University of Genova, Italy) and É. Schost have shown how working over the p-adics with only a few bits of precision makes it well-defined again. As an application, algorithms of better bit complexity are given for various parallel linear algebra problems over finite fields. A first version of this work was presented at the MEGA conference [21], and an extended article is in preparation.

5.3. Algorithms on sequences

Participants: Philippe Flajolet, Pierre Nicodème, Mireille Régnier, Bruno Salvy, Mathias Vandenbogaert.

Analytic combinatorics has allowed the team to solve numerous word or sequence problems: (i) one or several motifs, possibly infinite families, regular expressions, palindromes,... (ii) exact or degenerate motifs;

(iii) various probability models (Bernoulli, Markov, dynamic sources,...). Such analyses allow us to construct “toolkits" that make it possible to distinguish a significant signal from the noise, in several domains in computer science (text data, security systems, genomic data,...).

Our study of the distribution relies on the definition and manipulation of specific languages whose generating functions satisfy polynomial systems.

In a recent work, M. Régnier and A. Denise (Orsay University) studied the tail distributions for word occurrences. The combinatorial structure of the words allowed for the derivation of exact formulae of the rate function and an asymptotic expansion of the probabilities. A first application is the extraction of a weak signal hidden by a stronger signal. A second application is the assessment of the significance of clustered signals. This is the subject of the PhD thesis of E. Panina (NII Genetika), to be defended soon. These formulae have been implemented for the Markov model. The problem reduces to the solution of a polynomial equation, and have a low computational complexity. Our recent work [16] on public data from NIH (i) shows that these results are accurate and close to the results obtained by simulation or Monte-Carlo methods, when these methods are tractable; (ii) confirms theoretically that the normal approximation, widely used, is very poor; notably, it overestimates statistical significance.

Word counting procedures are implemented in C or Maple procedures. These results allowed M. Régnier and M. Vandenbogaert to participate in an international contest organized by M. Tompa (Washington University) between statistical softwares for InSilico prediction of regulatory signals [17].

In his thesis, defended in 2004, M. Vandenbogaert pointed out the noise introduced when the errors are uncontrolled and limited them to the ones allowed by the IUPAC code. In a collaboration with J. Clément (CNRS, Marne-la-Vallée University) and V. Boeva (Moscow University), M. Vandenbogaert and M. Régnier proposed a general definition of approximation that is consistent with the biological constraints on the so-called regulatory signals. Combinatorial formulae that allow for computing the waiting time for approximate words, either in the usual case or under this restriction, are given in Vandenbogaert’s thesis. An efficient algorithm, that mimicks the classical Aho-Corasick algorithm, has been designed to compute these formula [19]. A first application is given for fuzzy tandem repeats [4]. Tandem repeats are short repetitions that are hotspots for genome recombinations and are also related to some genetic diseases. This is implemented in the Tandem-SWAN software, available at http://strand.imb.ac.ru/swan/, that allows for the identification of weak clustered sites. These are needed for the analysis of several important regulatory systems such as nitrate/nitrite switch.

We collaborate on this subject with other INRIA projects. The algorithmic and combinatorial approach of ADAGE is complementary to our combinatorial and probabilistic approach, for instance on hidden words (see Flajolet’s work) or tandem repeats. Notably, the statistical computations that underly Tandem-SWAN algorithm can be reused by the software MREPS developed by G. Kucherov.

6. Contracts and Grants with Industry

6.1. Industrial Contracts

The Algorithms Project and Waterloo Maple Inc. (WMI) have developed a collaboration based on reciprocal interests. It is obviously interesting for the company to integrate functionalities at the forefront of the current research in computer algebra. Reciprocally, this integration makes our programs and our research visible to a very wide audience.

Numerous exchanges have thus taken place between the project and the company over the years. After more than 3 years within the project, J. Carette has been for several years Product Development Director at WMI, before going back to the academic world. Similarly, E. Murray, who worked for two years in the project developing the `combstruct` package is now working at WMI.

Thanks to all this activity, the company WMI considers INRIA as a special partner and grants it a free license for all of its research units. Moreover, a cooperation agreement has been signed between WMI and ALGO in 2001. In particular, one of the objectives is to replace all the routines dealing with asymptotic and series expansions in Maple by implementation of new algorithms dealing with very general classes of asymptotic scales.

7. Other Grants and Activities

7.1. National Actions

Aléa is a national working group dedicated to the analysis of algorithms and random combinatorial structures. It is a meeting place for mathematicians and computer scientists working in the area of discrete models. It is currently supported by CNRS (GDR A.L.P.) and is globally animated by Philippe Flajolet. In 2005, the yearly meeting (organized by C. Lavault) has gathered in Luminy over 80 participants from about 20 different research laboratories throughout France.

For the period 2003–2006, the Algo project participates in ACI-NIM a national research programme exploring New Interfaces of Mathematics. In this context, we take part in the ACPA project dedicated to paths and trees, probabilities and algorithms, this jointly with the Universities of Versailles, Bordeaux, and Nancy.

Since last year, a project called FLUX and involving the RAP project at INRIA as well as the University of Montpellier has been funded for a three year period by the national action ACI-MD relative to massive data: our objective is to develop high performance algorithms for the quantitative analysis of massive data flows an important problem in the monitoring of high speed computer networks.

For the period 2006–2009, the Algo project participates in a programme funded by the National Research Agency (ANR) entitled GECKO for “A Geometric Approach to Complexity and its Applications". Four teams are involved: ALGO (coordinator) and teams at the École polytechnique, the Universities of Toulouse and Nice. The project concentrates on three classes of objects: (i) univariate and multivariate polynomials (Newton process, factorization, elimination); (ii) structured matrices (whose coefficients can be polynomials); (iii) linear differential operators (noncommutative elimination, integration). The aim is to improve significantly the resolution of systems of algebraic or linear differential equations that appear in models, by taking geometry into account.

The National Research Agency (ANR) has funded this year a research project entitled SADA, whose goal is to investigate fundamental properties of random discrete structures and algorithms. The project duration is 3 years (Dec. 2005–Dec 2008). It involves five teams: ALGO/RAP from INRIA Rocquencourt, the Universities of Caen, Versailles, and Bordeaux (coordinator), as well as the Laboratory for Computer Science of the École polytechnique (LIX).

M. Régnier animates the project *Algorithmique et statistique des séquences* at the IMPG (*Informatique, Mathématique et Physique pour le Génome*).

7.2. Bilateral International Relations

Mireille Régnier is the French scientific head of a bioinformatics project supported by the French program ECO-NET, that involves three teams from Armenia, Georgia and Russia.

8. Dissemination

8.1. Animation

The ALGO project runs a biweekly seminar devoted to the analysis of algorithms and related topics. Several partner teams in the grand Paris area attend on a regular basis, and also take part in a yearly workshop, Aléa. Proceedings are collected and edited. This year’s publication [1] gathers 18 articles and one set of notes for a course given at Aléa’04.

Alin Bostan was a member of the poster committee of the conference ISSAC 2005.

Frédéric Chyzak has been a member of the program committee of this year's edition of the ISSAC conference, the premier international conference in computer algebra. He has served in Min WU's thesis committee, defended in Beijin (Academy of Mathematics and System Sciences, Academia Sinica).

Philippe Flajolet continues to serve as Chair of the Steering Committee of the international series of Conferences and Workshops called "*Analysis of Algorithms*". The 2005 edition was held in Barcelona and it attracted some 80 specialists of the area. He serves in a similar capacity as founder and chair of the French Working Group ALÉA supported by CNRS: the meeting was held at Luminy near Marseilles, and the participation neared 80 that year. Philippe Flajolet is also an external member of the Recruiting Committee (for computer science) at the École polytechnique. He is an editor of the journal *Random Structures and Algorithms*, an honorary editor of Theoretical Computer Science, and an honour member of the French association SPECIF. He also serves as one of the three editors of Cambridge University Press' prestigious series "Encyclopedia of Mathematics and its Applications". In 2005, he has served in thesis committees of several individuals: O. Gimenez (referee; Polytechnic University of Catalunya), X. Molinero (referee; Polytechnic University of Catalunya), S. Gerhold (referee; University of Linz), V. Puyhaubert (adviser; École polytechnique). He was also reviewer of several "Habilitation" memoirs, namely: F. Bassino (Paris-Marne la Vallée), G. Schaeffer (Bordeaux), E. Rivals (LIRMM, Montpellier). In 2005, Philippe Flajolet has served as member of the College of Reviewers for the Canada Research Chairs Program (mathematics and computer science), as well as an external reviewer for chairs (full professorships) at the Universities of Vienna and Turku. He remains a member of the French Academy of Sciences [15] and of the Academia Europaea. Finally, this year Philippe Flajolet has assumed the somewhat heavy responsability of chairing the Scientific Committee for Mathematics of the newly formed National Research Agency (ANR), which implied the heavy responsability of launching a programme of some 5 million Euros.

Éric Fusy gave talks at the Combinatorics Graphs and Computing colloquium in Berlin and at the Algorithms seminar of Caen on the subject of straight-line drawing of a triangulation. He gave talks at Bordeaux and Berlin (seminar of the Humboldt University) on the subject of a linear time algorithm for the random generation of planar graphs.

Mireille Régnier is a member of GTRI committee (International Relations) of INRIA COST. M. Régnier served in the program committee of a RECOMB satellite meeting on Regulation (2005). She co-organized the conference MCCMB'05 in Moscow. She was a member of the PhD committee of B. Bezhadi (École polytechnique).

Bruno Salvy is a member of the recruiting committees of the *Université des Sciences et Technologies de Lille* (in computer science) and of the University of La Rochelle (in mathematics). He is a member of the scientific counsil of the University of Versailles-St-Quentin. He is also member of the editorial board of the *Journal of Symbolic Computation* and of the *Journal of Algebra* (section Computational Algebra). This year, he has been engaged in a number of recruiting committees at INRIA Rocquencourt: junior researchers; promotion committee for research technicians (a category of administrative staff); he has organized the committees for post-docs and for the recruitment of researchers from other institutes and universities (*détachements* and *délégations*). He has also been in the scientific committee for the French national meeting on computer algebra and he is a member of the program committee of "Computational Geometry and Applications", a conference that will be held in Nice next June. He was a member of only one PhD committee this year, that of Benoît Daireaux (Caen)

8.2. Teaching

Alin Bostan, Frédéric Chyzak and Bruno Salvy together with Marc Giusti, François Ollivier and Éric Schost (the latter 3 are at École polytechnique) have set up and teach a course on computer algebra in the *Master Parisien de Recherche en Informatique* (MPRI).

Frédéric Chyzak, teaches several computer science courses as a *chargé d'enseignement à temps incomplet* at École polytechnique, including one in computer algebra.

Julien Fayolle teaches Maple (computer algebra system) in preparatory schools for the *Grandes Écoles* (CPGE-PCSI) at Fénelon High School, Paris.

Philippe Flajolet has been teaching about 20 hours in the fifth year programme (MPRI2) run jointly by Paris Universities and *Grandes Écoles*. In December 2004, he has been responsible for a complete three-day course on Probabilistic Algorithms in Chaing Mai (Thailand).

Frédéric Giroire has taught a “Networking" class and a “Performances Analysis" class in the Second year of IUP at the Paris VII university (equivalent Licence 3d year), and a “Project of Programmation" class in first year of Licence.

Carine Pivoteau teaches the Scheme programming language at the Paris VI University to first year Licence students.

Mireille Régnier teaches a 10 hours post-graduate course on “Combinatorics and Genome" at Évry and a 25 hours graduate course at the École Centrale de Paris, on the “Mathematical Problems and Algorithms in Genomics". She gave a few courses in new master “Bioinformatique et Biostatistiques " in Orsay. She is involved in the organization of the new option “Computer Science" to be opened in French Maths *agrégation* in 2006.

8.3. Participation in conferences, seminars, invitations

Alin Bostan has been invited for one week by J. von zur Gathen to visit the research group Algorithmic Mathematics in Paderborn, Germany. There he gave two talks on aspects related to Tellegen’s transposition theorem. He was an invited speaker at the workshop “Algebraic Complexity Theory meets Algorithmic Differentiation" organized at Humboldt Universität zu Berlin, Germany. Alin Bostan presented his work on the conversion from power sums to coefficients at the MEGA 2005 conference in Alghero, Italy and also during the *Journées nationales de calcul formel* at CIRM and at the Algorithms Project’s Seminar.

Frédéric Chyzak has presented joint work in progress with Ph. Dumas, H. Lê (CECM, Vancouver, Canada), J. Martins, M. Mishna (SFU, Vancouver, Canada), and B. Salvy in a talk about the desingularization of linear operators and their apparent singularities, at the *Journées nationales de calcul formel* (Luminy, France). He has also presented at the CECM (Simon Fraser University, Vancouver, Canada) his 2003 joint work with Moulay Barkatou and Michèle Loday-Richaud on algorithms for the extraction of multisections of a series solution of a given linear ODE.

Julien Fayolle presented his work on *Analyse de la profondeur dans un arbre des suffixes sous modèle markovien* at the *Journées Aléa*, Luminy. He gave a talk on “Typical Depth in Suffix Trees" at the 2005 Conference on the Analysis of Algorithms, Barcelona, Catalunya.

Philippe Flajolet has been invited for a colloquium at the Free University of Berlin. His other invitations abroad include the Universities of Barcelona and Linz. He has been invited to give a talk at the Mathematical Physics seminar of the University Paris VII (on “Airy phenomena in analytic combinatorics”) as well as at the Algebra and Geometry Seminar at the University of Versailles (on “Algebraic aspects of analytic urns”). He has also been one of the invited speakers at *Séminaire Lotharingien de Combinatoire*, which despite its name is an international event (lecture on “The Fermat cubic, elliptic functions, continued fractions, and a combinatorial excursion”).

Frédéric Giroire presented his work on “Order Statistics and Estimating Cardinalities of Massive Datasets” at the Analysis of Algorithms conference (Barcelona, Spain).

Micha Hofri gave a presentation on “Analysis of Approximate Median Selection" at the Algorithms Project’s Seminar.

Carine Pivoteau presented joint work with Philippe Flajolet and Éric Fusy on “Boltzmann random generation” at the Algorithms Project’s Seminar, at the Calfor Seminar (LIP6) and at the *Journées Arbres Ou Cartes* in Bordeaux.

Mireille Régnier presented her results at Lille, LIX workshop 2005, the conference MCCMB’05 in Moscow and the RECOMB meeting on Regulation in San Diego, USA.

Bruno Salvy has been one of the three invited speakers (together with Bruno Buchberger and Wen-Tsun Wu) at ISSAC'05 (Beijing), the premier international conference in computer algebra, where he gave a presentation on “D-finiteness: Algorithms and Applications”. He has also been invited in Banff (Canada) for a workshop on “Challenges in Linear and Polynomial Algebra in Symbolic Computation Software”.

8.4. Foreign Visitors

A large number of our visitors have given talks at the seminar of the project. This year, we received:

Mathias Vandenbogaert (BioZentrum, University of Basel, Switzerland); Bernard Chazelle (Princeton University, USA); Robert Sedgewick (Princeton University, USA); Alfredo Viola (INCO-Montevideo, Uruguay); Wojtek Szpankowski (Purdue University, USA); Stefan Gerhold (RISC, University of Linz, Austria); Gerald Kok (TU Delft, The Netherlands); Nicolas Broutin (McGill University, Montreal, Canada).

9. Bibliography

Books and Monographs

[1] F. CHYZAK (editor). *Algorithms Seminar, 2002–2004*, Research Report, 120 pages, vol. 5542, Institut National de Recherche en Informatique et en Automatique, April 2005, http://www.inria.fr/rrrt/rr-5542.html.

[2] P. FLAJOLET, R. SEDGEWICK. *Analytic Combinatorics*, Chapters I–IX of a book to be published, 688p.+x, October 2005, http://algo.inria.fr/flajolet/Publications/book051001.pdf.

Doctoral dissertations and Habilitation theses

[3] V. PUYHAUBERT. *Modèles d'urnes et phénomènes de seuils en combinatoire analytique*, Ph. D. Thesis, École polytechnique, 2005.

Articles in refereed journals and book chapters

[4] V. BOEVA, V. MAKEEV, D. PAPATSENKO, M. RÉGNIER. *Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression*, in "Bioinformatics", 10 pages., To appear.

[5] A. BOSTAN, P. FLAJOLET, B. SALVY, É. SCHOST. *Fast computation of special resultants*, in "Journal of Symbolic Computation", vol. 41, n^o 1, January 2006, p. 1–29.

[6] A. BOSTAN, É. SCHOST. *Polynomial evaluation and interpolation on special sets of points*, in "Journal of Complexity", Festschrift for the 70th Birthday of Arnold Schönhage, vol. 21, n^o 4, August 2005, p. 420–446.

[7] F. CHYZAK, M. MISHNA, B. SALVY. *Effective Scalar Products of D-finite Symmetric Functions*, in "Journal of Combinatorial Theory, Series A", vol. 112, n^o 1, oct 2005, p. 1–43.

[8] F. CHYZAK, A. QUADRAT, D. ROBERTZ. *Effective algorithms for parametrizing linear control systems over Ore algebras*, in "Applicable Algebra in Engineering, Communication and Computing", 58 pages. In press, 2005, http://dx.doi.org/10.1007/s00200-005-0188-6.

[9] J. A. FILL, P. FLAJOLET, N. KAPUR. *Singularity Analysis, Hadamard Products, and Tree Recurrences*, in "Journal of Computational and Applied Mathematics", vol. 174, February 2005, p. 271–313.

[10] P. FLAJOLET, S. GERHOLD, B. SALVY. *On the non-holonomic character of logarithms, powers, and the nth prime function*, in "The Electronic Journal of Combinatorics", A2, 16 pages, vol. 11, n^o 2, April 2005.

[11] P. FLAJOLET, M. NEBEL, H. PRODINGER. *The scientific works of Rainer Kemp (1949–2004)*, in "Theoretical Computer Science", In press. 15 pages.

[12] P. FLAJOLET, J. GABARRÓ, H. PEKARI. *Analytic Urns*, in "Annals of Probability", Available from ArXiv:math.PR/0407098., vol. 33, n^o 3, 2005, p. 1200–1233.

[13] M. GIUSTI, G. LECERF, B. SALVY, J.-C. YAKOUBSOHN. *On Location and Approximation of Clusters of Zeroes: Case of Embedding Dimension One*, in "Foundations of Computational Mathematics", 51 pages. Accepted, November 2005, To appear.

[14] M. GIUSTI, G. LECERF, B. SALVY, J.-C. YAKOUBSOHN. *On Location and Approximation of Clusters of Zeroes of Analytic Functions*, in "Foundations of Computational Mathematics", 45 pages, 2005.

[15] GÉRARD. HUET, P. FLAJOLET. *Mathématiques et Informatique*, J.-C. YOCCOZ (editor). , Rapport sur la science et la technologie, Académie des sciences, vol. 20, TEC&DOC, Paris, 2005.

[16] M. RÉGNIER, M. VANDENBOGAERT. *Comparison of Statistical Significance Criteria*, in "Journal of Bioinformatics and Computational Biology", 12 pages. In press, To appear.

[17] M. TOMPA, N. LI, T. BAILEY, G. CHURCH, B. DE MOOR, E. ESKIN, A. FAVOROV, M. FRITH, Y. FU, J. KENT, V. MAKEEV, A. MIRONOV, W. NOBLE, G. PAVESI, G. PESOLE, M. RÉGNIER, N. SIMONIS, S. SINHA, G. THIJS, J. VAN HELDEN, M. VANDENBOGAERT, Z. WENG, C. WORKMAN, C. YE, Z. ZHU. *An Assessment of Computational Tools for the Discovery of Transcription Factor Binding Sites*, in "Nature Biotechnology", vol. 23, n^o 1, January 2005, p. 137–144.

Publications in Conferences and Workshops

[18] M. BARDET, J.-C. FAUGÈRE, B. SALVY, B.-Y. YANG. *Asymptotic Behaviour of the Degree of Regularity of Semi-Regular Polynomial Systems*, in "MEGA'05", Eighth International Symposium on Effective Methods in Algebraic Geometry, Porto Conte, Alghero, Sardinia (Italy), May 27th – June 1st, 2005.

[19] V. BOEVA, J. CLÉMENT, M. RÉGNIER, M. VANDENBOGAERT. *Assessing the significance of Sets of Words*, in "Combinatorial Pattern Matching 05", Lecture Notes in Computer Science, In Proceedings CPM'05, Jeju Island, Korea, vol. 3537, Springer Verlag, 2005, p. 358–370.

[20] A. BOSTAN, T. CLUZEAU, B. SALVY. *Fast Algorithms for Polynomial Solutions of Linear Differential Equations*, in "ISSAC'05, New York", M. KAUERS (editor). , Proceedings of the 2005 International Symposium on Symbolic and Algebraic Computation, July 2005, Beijing, China., ACM Press, 2005, p. 45–52.

[21] A. BOSTAN, L. GONZÁLEZ-VEGA, H. PERDRY, É. SCHOST. *From Newton sums to coefficients: complexity issues in characteristic p*, in "MEGA'05", Eighth International Symposium on Effective Methods in Algebraic Geometry, Porto Conte, Alghero, Sardinia (Italy), May 27th – June 1st, 2005.

[22] J. FAYOLLE, M. D. WARD. *Analysis of the Average Depth in a Suffix Tree under a Markov Model*,

in "2005 International Conference on Analysis of Algorithms", C. MARTÍNEZ (editor). , DMTCS Proceedings, vol. AD, Discrete Mathematics and Theoretical Computer Science, 2005, p. 95–104, http://www.dmtcs.org/proceedings/html/dmAD0109.abs.html.

[23] É. FUSY. *Transversal structures on triangulations, with application to straight line drawing*, in "Graph Drawing 2005", Lecture Notes in Computer Science, Springer-Verlag, To appear.

[24] É. FUSY. *Quadratic exact size and linear approximate size random generation of planar graphs*, in "2005 International Conference on Analysis of Algorithms", C. MARTÍNEZ (editor). , DMTCS Proceedings, vol. AD, Discrete Mathematics and Theoretical Computer Science, 2005, p. 125-138, http://www.dmtcs.org/proceedings/abstracts/dmAD0112.abs.html.

[25] É. FUSY, D. POULALHON, G. SCHAEFFER. *Dissections and trees, with applications to optimal mesh encoding and to random sampling*, in "SODA", Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005, Vancouver, British Columbia, Canada, January 23-25, 2005, 2005, p. 690–699.

[26] F. GIROIRE. *Order statistics and estimating cardinalities of massive data sets*, in "2005 International Conference on Analysis of Algorithms", C. MARTÍNEZ (editor). , DMTCS Proceedings, vol. AD, Discrete Mathematics and Theoretical Computer Science, 2005, p. 157-166, http://www.dmtcs.org/proceedings/html/dmAD0115.abs.html.

[27] B. SALVY. *D-Finiteness: Algorithms and Applications*, in "ISSAC'05", M. KAUERS (editor). , Abstract for an invited talk. Proceedings of the 2005 International Symposium on Symbolic and Algebraic Computation, Beijing, July 2005., ACM Press, 2005, p. 2–3.

Miscellaneous

[28] E. V. F. CONRAD, P. FLAJOLET. *The Fermat cubic, elliptic functions, continued fractions, and a combinatorial excursion*, Submitted to Séminaire Lotharingien de Combinatoire, 44 pages, July 2005.

[29] C. PIVOTEAU. *Génération aléatoire et modèles de Boltzmann*, Mémoire de DEA, Université Paris VI, 2005.

[30] A. VERA. *Analyse dynamique d'algorithmes Euclidiens en dimension 2*, Mémoire de DEA, Université Paris VI, 2005.