INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Algo

Algorithms

Paris - Rocquencourt

THEME SYM

Activity Report

2007

Table of contents

1. Team

Head of project-team
Bruno Salvy [Research Director (DR) Inria]

Vice-head of project team
Philippe Flajolet [Research Director (DR) Inria, HdR]

Administrative assistant
Virginie Collette [(TR) Inria]

Research scientists (Inria)
Alin Bostan [Research Associate (CR) Inria]
Frédéric Chyzak [Research Associate (CR) Inria]
Mireille Régnier [Research Director (DR) Inria, HdR]

Research scientists (partners)
Philippe Dumas [Professor, Cl. Prépa. lycée Jean-Baptiste Say]

Ph. D. students
Shaoshi Chen [starting December]
Éric Fusy [Détachement du Corps des Télécoms, until September]
Carine Pivoteau [University of Paris VI]

PostDoc
Valentina Boeva [University of Moscow, from April to September]
Nicolas Broutin [McGill University, from September]
Frédéric Meunier [Détachement du Corps des Ponts et Chaussées, until August]
Nicolas Le Roux [Professeur en disponibilité, Rectorat de Créteil, from September]

Students intern
Pranav Kashyap [Indian Institute of Technology, Numbai, from May to July]
Marc Mezzarobba [ENS Paris, from February to July]

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 focuses 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 analyze, 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 into the standard library of Maple, but the most up-to-date version is always available for free from our web pages. 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 have been the standard way to solve polynomial systems in Maple for a long period (although the user would 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 encyclopedias 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 behavior have been computed automatically. It gets more than 16,000 hits per month. 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 role of a showcase for part of the packages developed in our project. It gets 27,000 hits per month.

A new package, *MultiSeries* has been developed recently. It implements so-called multi-series, that are series in general asymptotic scales, each of whose coefficient is itself potentially a new series. This makes it possible to handle in a transparent and dynamic way the problems of finding the proper asymptotic scale for an expansion and of dealing with the indefinite cancellation problem. This package is designed in such a way that it can take the place of the existing *series*, *asympt* and *limit* Maple functions, in a totally transparent manner.

Tandem repeats are short repetitions that are hotspots for genome recombinations and are also related to some genetic diseases. *TandemSWAN* searches for degenerate tandem repeats without insertions and deletions, but with a high substitution rate. It is based on calculation of the repeat statistical significance and identifies the length of the repeated unit and the number of repetitions. It allows for the identification of weak clustered sites, that are needed for the analysis of several important regulatory systems such as nitrate/nitrite switch. It is written in C language, using some C++ features.

Software *AhoPro* allows for counting various probabilities for exceptional words. It relies on automata theory and an updated version of classical Aho-Corasick algorithm. It improves the computation of existing functions. It takes into account new problems, such as motifs cooccurrences and complex clusters. It is written in C and available on line (http://bioinform.genetika.ru/AhoPro/). It is integrated to SeSiMCMC software of associate tean GosNIIGenetikam that searches Transcription Factor Binding Sites on Genomes.

5. New Results

5.1. Analysis of algorithms

Participants: Nicolas Broutin, Frédéric Chyzak, Philippe Flajolet, Éric Fusy, Pranav Kashyap, Frédéric Meunier, Mireille Régnier, Bruno Salvy.

The reference book on analytic combinatorics [1], by Flajolet and Sedgewick, is nearing completion: it has currently 803 pages (a free web edition is available) and will be published in 2008 by Cambridge University Press. It aims to provide a broad and accessible coverage of the interactions between discrete and continuous mathematics, in the perspective of detailed analyses of combinatorial models, as may be present in the applied sciences as well as in analysis of algorithms. The core theory of analytic combinatorics, as presented in [1], is making advances along three major axes: (i) symbolic methods; (ii) complex asymptotics; (iii) random structures and probabilities. Point (i) addresses the issue of setting up equations that automatically translate a given combinatorial-probabilistic model. Point (ii) aims at developing general-purpose methods for extracting asymptotics on coefficients of generating functions, a task best placed within the framework of complex analysis. Point (iii) is evolving in the direction of classifying the most important processes of combinatorics, in a way that often provides an attractive alternative to stochastic theory. A global survey of analytic combinatorics appears in [22], as a plenary invited lecture at ACM-SIAM SODA (Symposium on Discrete Algorithms).

In [8], we pursue the asymptotic study of divide-and-conquer sequences. We deal with a very specific type of such sequences, namely the translations via a radix numeration system of rational power series in non-commutative indeterminates, which occur in formal linguistics. As a consequence, we develop a systematic approach, which permits us to deal with a real-life model of large size arising from computational number theory and cryptography. This prepares a general study of this type of sequences by complex-analytic methods: the aim is to obtain a general theorem classifying their asymptotic behaviour, while at this time only examples of small size have been studied. Some perhaps surprising applications of complex-analytic methods to difference calculus appear in [9].

An unexpected outcome of the general theory of analytic combinatorics is in the *random generation* of large structured combinatorial configurations. Owing to an original confluence of ideas between statistical physics (Boltzmann models), analytic combinatorics (symbolic approaches), and computer algebra (the systematic evaluation of generating functions near singularities), we can now foresee the existence of a highly general framework in which it should be possible to express large classes of models of interest in various branches of computer science (from constrained sequences to trees, structured graphs, and automata); then, the design and the production of extremely efficient simulation algorithms can be automated. The application of these ideas to a large (infinite) family of models, leading to highly efficient simulation algorithms form sthe subject of [24]. A striking application [29] of the idea is a highly efficient algorithm for the random generation of planar graphs: the complexity is quadratic for exact size generation and even *linear*, as soon as a small tolerance on size is allowed. Another random generator, for plane partitions, is given in [27]. These partitions are extensively studied in statistical physics. The generator is highly efficient, being quasi-linear, in order to generate a plane partition under the Boltzmann distribution.

Éric Fusy has published several results giving explicit bijective (i.e., one-to-one) constructions for planar structures such as polygonal meshes and more generally planar maps (graphs embedded on a planar surface) [10], [11], [12]. These results, which are currently in press and are thoroughly described in the thesis of Éric Fusy [2], have important applications. Indeed, such bijective constructions yield succinct encoding procedures relevant in the context of mesh compression, and linear time random generation procedures. The article [12] provides a linear time random sampler for polygonal meshes of the sphere. The paper on the enumeration of unlabeled planar graphs has now be published [3]. This is a promising step toward the enumeration of unlabeled planar graphs, a famously unsolved problem. A further progress has been realised recently by Éric Fusy, who describes a systematic method to enumerate families of graphs from their 3-connected components in [30].

Complementary studies in the area of structural combinatorics have been conducted by F. Meunier, during his postdoctoral stay. A paper [15] gives the first combinatorial proof of a generalization of Sperner's lemma found by R. Freund. The existence of this proof was an open question. A parity assertion similar to that of Sperner's lemma, in connexion with this question, had also been conjectured; the paper settles both questions. Also a 31 page chapter in a collective work relative to graph theory has been competed [16]. This chapter offers a contemporary point of view on algorithmic aspects of cuts and walks in graphs, presented at a graduate level. Advanced topics, like the link between the chinese postman problem and integers multiflows, are also treated. Since it deals with algorithmic questions, this work is directly relevant to the global aims of the ALGO project.

Regarding algorithms for massive data sets and data mining, we have designed this year the HyperLogLog counting algorithm, which holds the world record for estimatig *cardinalities* (number of distinct elements) of large data ensembles. It is important to note that the very design of the algorithm is based on a subtle exploitation of fine characteristics of the underlying discrete model. With a few thousand bits of memory, this algorithm makes it possible to estimate the number of distinct elements in a stream of data of several tens of gigabytes, with an accuracy of few percent (typically 1.5% with 1024 bytes). In this context, no a priori probabilistic assumption is made on the nature of data: the algorithms are universal. These algorithms have applications in the area of data mining and especially network monitoring. Pranav Kashyap, from IIT Bombay, has developed a comprehensive library and thoroughly validated a large number of derived algorithms in the area of document comparison and elephant/mice detection. An efficient "sliding window" version of some of these algorithms has also been finally developed.

5.2. Computer Algebra

Participants: Alin Bostan, Frédéric Chyzak, Philippe Flajolet, Nicolas Le Roux, Marc Mezzarobba, Bruno Salvy.

High precision expansions of power series solutions of differential equations are needed in various branches of computational mathematics, from combinatorics, where the desired power series is a generating function, to numerical analysis and computational number theory. In [19], A. Bostan, F. Chyzak, B. Salvy, together with F. Ollivier (École Polytechnique), É. Schost (U. Western Ontario, Canada) and A. Sedoglavic (U. Lille), give fast algorithms for computing many coefficients of power series solutions of systems of differential equations. The new algorithms use a number of arithmetic operations which is quasi-linear in the number of computed terms. Moreover, these algorithms are optimal, in the sense that the cost of resolution is proportional to that of checking the solution.

In [6], A. Bostan, F. Morain (project-team Tanc), B. Salvy and É. Schost applied the previous result to a problem in elliptic curve cryptography. They introduced an algorithm that computes a special rational morphism between two elliptic curves; its complexity is quasi-linear with respect to the degree of the morphism. The new algorithm allows to speed-up a crucial step of the Schoof-Elkies-Atkin algorithm for determining the cardinality of an elliptic curve over a finite field.

The case of linear differential equations with polynomial (rather than power series) coefficients allows for specific algorithms and the Algo project-team has been spending a large part of its effort on these for many years.

A. Bostan, P. Gaudry (project-team Cacao) and É. Schost study the complexity of computing one or several terms in a linear recurrence with polynomial coefficients in [5]. Using the classical "baby step / giant step" technique and new results on polynomial extrapolation, they improve an old algorithm due to the Chudnovsky brothers. As an application, they obtain the best currently known upper bounds for factoring integers deterministically. A second application is to computing the Cartier-Manin operator, which is useful for point counting on hyperelliptic curves in large characteristic.

The analytic solutions of these differential equations can be evaluated efficiently up to very high precisions. In [32], M. Mezzarobba describes what seems to be the first general implementation of this process. He also suggests a few improvements to the relevant algorithms, especially concerning the computation of sharp upper bounds for the tails of power series expansions of those functions.

It is classical that algebraic series are solutions of linear differential equations. In [18], A. Bostan, F. Chyzak, G. Lecerf, B. Salvy and É. Schost have given a fast algorithm for the expansion of these series. The algorithm is based on this classical idea and on a precise analysis of the total size of several equations satisfied by algebraic series and their coefficient sequences (respectively, differential equation of minimal order, recurrence of minimal order, and of minimal size). Interestingly, this work has revealed a new approach that should generalize and help computing symbolic integrals and sums without suffering from the growth of intermediate objects.

A follow-up of methods for linear differential equation is an application to linear control systems. A collaboration of F. Chyzak with A. Quadrat (project-team Café) and D. Robertz (U. 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 developed [21], based on Chyzak's Maple implementation of Gröbner tools.

For several years, B. Salvy and A. Bostan have been working jointly with the LIX 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, M. Giusti (École polytechnique) and J.-C. Yakoubsohn (U. 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 polynomial systems [13] under a technical condition of "embedding dimension 1".

Structured linear algebra techniques are a versatile set of tools, which enable to deal at once with various types of matrices, with features such as Toeplitz-, Hankel-, Vandermonde- or Cauchy-likeness. It is classical that such linear systems can be solved, by means of a compact representation, in $\tilde{O}(\alpha^2 n)$ operations, where n is the matrix size and α is a measure of the structure. Together with Cl.-P. Jeannerod (project-team Arenaire) and É. Schost, A. Bostan showed that this cost can be reduced to $\tilde{O}(\alpha^{\omega-1} n)$, where $\omega < 2.38$ is a feasible exponent for matrix multiplication. The improvement is based on re-introducing fast dense linear algebra into operations involving the compact generators of the given matrix. This has consequences for efficient Hermite-Padé approximation and interpolation of bivariate polynomials.

5.3. Algorithms on sequences

Participants: Valentina Boeva, Philippe Flajolet, Mireille Régnier, Bruno Salvy.

Our goal is to combine analytic combinatorics to establish general mathematical results and algorithms that provide efficient optimized implementations for specific data constraints.

Our group studies combinatorial properties of words in order to improve searching algorithms and extract unexpectedly frequent or rare motifs from random sequences. Our main motivation is to relate such motifs to biological functions on the genome. Ultimately, statistical or probabilistic results are to be integrated into string searching algorithms or software.

Starting from exact counting formulae derived by the tools of analytic combinatorics, we focus on efficient implementations and tight approximation derivations. On the one hand, efficient implementations have been realized for sets of words, that rely on automata theory. Our main application is the study of words, also called sites on the genome, that may be recognized by different proteins, in a competing or synergistic manner. On the other hand, approximations have been derived, with a tight error bound. Recent results with A. Denise (Orsay University) allow to compute the probability of rare events, commonly called *p-value* by bioinformaticians, with a linear algorithm. This is a drastic improvement on the existing exponential algorithm. Tightness of these *p-value* expressions allows to compute conditional probabilities, and extract weak signals hidden by stronger signals. *P-value* computation was implemented in software *SSiMCMC* developed at NIIGenetika, with whom we collaborate in associate team MIGEC. As a whole, systematic comparisons of commonly used statistical criteria for exceptional words, and precise statements on their validity domains, have been realized with M. Vandenbogaert. In his master's thesis in 2006, B. Besson studied the mathematical grounds of the so-called Position Specific Scoring Matrix model, PSSM, in favour among biologists. Notably, he established a link between the number of available experimental data used to build the model, and the precision and significance of the model obtained. This work was continued in 2007 in a collaboration with J. S. Varré (SEQUOIA, Lille). A special attention has been paid to Tandem Repeats. Tandem repeats are short repetitions that are hotspots for genome recombinations and are also related to some genetic diseases. *TandemSWAN* It allows for the identification of weak clustered sites, that are needed for the analysis of several important regulatory systems such as nitrate/nitrite switch. The originality of the software *TandemSWan*, developed jointly in C with NIIGenetika, is the implementation of a procedure to assess the statistical significance of repeats, that can be reused in other softwares such as *Mreps* developed by SEQUOIA, Lille, by G. Kucherov. Counting sites clusters, or assessment of a statistical significance for co-occurring motifs had not be addressed so far. Our algorithm *AhoPro* [4] relies on automata theory and an improvement of classical AhoCorasick algorithm.

RNA secondary structures prediction is a hard problem, as a given sequence has an exponential number of potential foldings. It is even worse when pseudo-knots are taken into account. Finding a structure, e.g., a set of helices, that is common to sequences from phylogenetically closed organisms turned out to be the most efficient way. This comparative approach is meaningful as sequencing is often realized now for closely related genomes. Our heuristic relies on a few simple probability criteria to chose common helices, and led recently to a variant, *P-DCfold*, that deals with pseudo-knots. This work is currently continued by M. Ganjtabesh, PhD student at LIX, that extends our counting results when pseudo-knots are allowed (joint work with P. Nicodème).

In her thesis cosupervised by M. Régnier, G. Boldina (Almaty University) studies specific motifs that allow for splicing introns. Roughly speaking, the DNA sequence between splicing sites is a non-coding sequence that does not participate to the translation of DNA into a protein sequence, and therefore should be excised. It turns out that for U2-type introns, that represent approximately 99,6% of introns for eucaryotes, these motifs are highly degenerated; therefore, they do not resort to usual methods. A definition of a hydropathy profile allows to capture a biochemical property, and it has been shown that splice sites, and intron types as well, are associated to abnormal profiles [31]. The statistical assessment of this abnormality relies on large deviations results being currently developed in the project-team ALGO.

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 and has continued till April of this year.

Starting this Fall, our work on automating the derivation of formulæ for special functions takes place in the joint Inria-Microsoft Research Lab.

7. Other Grants and Activities

7.1. National Actions

Alea 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 March 2007, the yearly meeting (organized by Frédérique Bassino et Julien Clément) has gathered in Luminy over 80 participants from about 20 different research laboratories throughout France.

For the period 2003–2007, 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 2005 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 also funded 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).

7.2. Bilateral International Relations

Mireille Régnier is the French scientific head of Inria associated team MIGEC with Russian partner GosNI-IGenetika. She is also the head of a bioinformatics project supported by the French program ECO-NET, that involves three teams from Armenia, Kazakhstan and Russia.

She is a participant of IST INTAS: This is a grant to work on comparative genomics of bacteria. This project is joint with TUM (Munchen), CNR-ITB (Milan), NII-Genetika (Moscow) and Moscow University.

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, Alea. Proceedings are collected and edited.

Frédéric Chyzak has been a co-organizer of the 2007 edition of the French national meeting in computer algebra (JNCF’07), that has gathered some 80 participants in Luminy. He still is co-organizer of the 2008 edition (JNCF’08). He is a member of the recruiting committee of Univ. Limoges, in mathematics. He has presented the work on algebraic series [18] during his visit to Z. Li’s team in Beijing.

Éric Fusy has been a member of the progam committee of the Workshop on Analytic Algorithms and Combinatorics ANALCO’08.

Philippe Flajolet 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”.

He serves as Chair of the Steering Committee of the international series of Conferences and Workshops called “Analysis of Algorithms”. The yearly edition attracts some 80 specialists of the area. He serves in a similar capacity as founder and chair of the French Working Group Alea supported by the GDR-IM [“mathematical informatics”] of CNRS: the yearly meetings are held at Luminy near Marseilles, and the participation nears 80 every year.

Philippe Flajolet is also an external member of the Recruiting Committee for computer science at the École polytechnique. He is a member of the board of experts for the Canada Research Chairs.

Mireille Régnier has coorganized the bioinformatics conference MCCMB'07 in Moscow. She participated to the organization of the new computer science option at the French *agrégation* of mathematics. She served in the committee in 2007. She was a member of the recruiting committees of the University of Lille (in computer science). She has been a reviewer for the PhD C, Herrbach (LRI, Orsay). She was a member of COST-GTRIm for international relationships of Inria.

Bruno Salvy has been on the program committee of the conference "Analysis of Algorithms", Nice, 2007 and he is on the program committee for the conference "Formal Power Series and Algebraic Combinatorics", Valparaiso, Chile, 2008. He is organizing the working group Computer Algebra of the CNRS GDR IM ("Mathematical Informatics"). He is a member of the editorial board of the Journal of Symbolic Computation and of the Journal of Algebra (section Computational Algebra).

He is a member of the recruiting committees of the University of Lille (in computer science), of the University of La Rochelle (in mathematics). At Inria Rocquencourt, he has organized the committees for post-docs and for the recruitment of researchers from other institutes and universities (détachements and délégations) together with Stéphane Gaubert.

This year, he has been a member of the PhD committee of P. Rémy (Mathematics, U. Angers).

8.2. Teaching

Alin Bostan, *Frédéric Chyzak*, and *Bruno Salvy* have set up and taught a 48h course in computer algebra together with Marc Giusti (from École polytechnique). This course is part of the *Master Parisien de Recherche en Informatique* (MPRI).

Frédéric Chyzak teaches a course in computer algebra as a *chargé d'enseignement à temps incomplet* at École polytechnique.

This year, *Alin Bostan* and *Bruno Salvy* have also set up and taught a new 48h course at the *École Normale Supérieure* on computer algebra oriented towards experimental mathematics.

Alin Bostan, *Frédéric Chyzak*, and *Bruno Salvy* gave in March 2007 a 4-hour course on "D-finiteness: Algorithms and Applications" at the École Jeunes Chercheurs en Informatique Mathématique (EJCIM'07), in Nancy.

Philippe Flajolet has given a 12 hour course at the Parisian Master of Research in Computer Science (MPRI), on the analysis of algorithms. He has also taught intensive courses on analytic combinatorics in Santiago de Chile and Barcelona (CRM). In this last course, *Éric Fusy* has been a teaching assistant of Philippe Flajolet; he was co-organising the exercice sessions (5 sessions of 2 hours) with Omer Giménez.

Mireille Régnier teaches a 12/year of courses in the Bioinformatics Master of Évry on combinatorics and algorithms in genomics. She was also invited to give two weeks of postgraduate courses in Almaty (Kazakhstan).

8.3. Participation in conferences, seminars, invitations

Alin Bostan, *Frédéric Chyzak*, and *Bruno Salvy*'s joint work with Grégoire Lecerf and Éric Schost [18] has been presented at the conference ISSAC'07 (Waterloo, Ontario, Canada), July 2007. *Alin Bostan*'s joint work with Claude-Pierre Jeannerod and Éric Schost [20] has also been presented there. He gave a talk entitled "Solving Toeplitz- and Vandermonde-like Linear Systems with Large Displacement Rank", in November 2007 in Nice, for the yearly workshop of the Gecko project of the ANR.

Frédéric Chyzak has presented preliminary work joint with A. Bostan and B. Salvy of the project-team and Z. Li (Academy of Mathematics and System Sciences, Beijing) at the yearly workshop of the Gecko project of the ANR. Expected results are fast algorithms for the computation of least common multiples of linear differential and difference operators.

Philippe Flajolet has been an invited plenary speaker at the joint ACM–SIAM Symposium on Discrete Algorithms (SODA'07, New Orleans) and the "KnuthFest" (Bordeaux, October 2007), and has given a main tutorial/survey talk at the international conference Analysis of Algorithms (AofA'07, Juan-les-Pins, June 2007). He has been also invited speaker at the British Combinatorics Day (Oxford, February 2007) Workshop on Probabilistic and Algebraic Methods in Combinatorics (CRM, Barcelona, June 2007). He has given seminars and colloquia at the Universities of Grenoble (number theory, February 2007), Lyon (Computer Science Colloquium, ENS, April 2007), Orleans (Mathematics Colloquium, September 2007), and Paris-Sud at Orsay (Computer Science Colloquium, December 2007).

Éric Fusy has given an invited talk at the workshop on combinatorial problems arising from statistical physics at Montreal (Feb. 2007). He has been invited for one week at the university of Oxford (May 2007) where he colaborated with Bilyana Shoilekova on graph enumeration problems and gave a talk on random generation of planar graphs. He gave one invited talk in Barcelona at the Conference on Enumeration and Probabilistc Methods in Combinatorics; he also gave two talks in Berlin in July, in the discrete mathematics group at Technische Universitaet Berlin and in the Algorithmics group at Humboldt Universitaet zu Berlin. He was invited for one week in Zurich in the theoretical computer science group where he colaborated with Konstantinos Panagiotou on statistical properties of planar graphs and gave an invited talk on planar drawing of triangulations. He was invited by the combinatorics group at Labri, Bordeaux, where he colaborated with Mireille Bousquet-Mélou and Nicolas Bonichon, and gave a seminar on bijective constructions for planar maps. He gave an invited talk at the AMS Special Session on Algorithmic Probability and Combinatorics, held at DePaul University Chicago (Oct. 2007).

Nicolas Le Roux gave a talk about the computation of formal local series solutions of linear partial differential equations systems at the meeting of the Gecko project of the ANR held in INRIA Sophia Antipolis (november 2007).

Frédéric Meunier gave a talk at the SMAI 2007 conference on a combinatorial optimization problem – the so-called paint shop problem. At the AofA'07 conference in Juan-les-Pins, he presented his joint work [23] that deals with a new on-line algorithm dedicated to estimating the number of distinct elements in very large data sets.

Mireille Régnier visited three times the associated team NIIGenetika in Moscow. She gave a talk at INRIA-NIH workshop in Washington, DC and at workshop AlgoBio in Moscow. She was invited by scientific council of Nantes University for a presentation on Genomics.

Bruno Salvy has been invited to give a talk at the Joint Mathematics Meeting of the American Mathematical Society and the Mathematical Association of America, on *Gfun: 20 years later* in New Orleans. He presented [19] at the ACM-SIAM symposium on Discrete Algorithms (SODA), also in New Orleans. He gave a talk on algebraic series following [18] at the yearly workshop of the Gecko project of the ANR.

8.4. Foreign Visitors

A large number of our visitors have given talks at the seminar of the project. This year, we received: Bilyana Shoilekova, Oxford University, U.K.; Michel Schellekens, National University of Ireland, Cork, Ireland Alfredo Viola, University of Montevideo, Uruguay; Ziming Li, Academy of Mathematics and System Sciences, Beijing, China; Manuel Kauers, Johanes Kepler University, Linz Austria; Vsevolod Makeev, Laboratory of Bioinformatics, GosNIIgenetika, Moscow, Russia; Julia Medvedeva, Russian Academy of Sciences, Moscow, Russia; Ivan V. Kulakovsky, Russian Academy of Sciences, Moscow, Russia; Zara Kirakossian, Yerevan State University, Yerevan, Republic of Armenia; Mikhail Roytberg, Institute of Mathematical Problems in Biology, Puschino, Russia.

9. Bibliography

Year Publications

Books and Monographs

[1] P. FLAJOLET, R. SEDGEWICK. *Analytic Combinatorics*, Chapters I–IX of a book to be published by Cambridge University Press, 791p.+xii, available electronically from P. Flajolet's home page, 2007, http://algo.inria.fr/flajolet/Publications/.

Doctoral dissertations and Habilitation theses

[2] É. FUSY. *Combinatoire des cartes planaires et applications algorithmiques*, Ph. D. Thesis, École Polytechnique, June 2007.

Articles in refereed journals and book chapters

[3] M. BODIRSKY, É. FUSY, M. KANG, S. VIGERSKE. *Enumeration and Asymptotic Properties of Unlabeled Outerplanar Graphs*, in "Electronic Journal of Combinatorics", 24 pages, vol. 14, n^{o} R66, September 2007.

[4] V. BOEVA, J. CLÉMENT, M. RÉGNIER, M. ROYTBERG, V. MAKEEV. *Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules*, in "Algorithms for molecular biology", vol. 2, n^{o} 13, 2007, 25 pages.

[5] A. BOSTAN, P. GAUDRY, É. SCHOST. *Linear recurrences with polynomial coefficients and application to integer factorization and Cartier-Manin operator*, in "SIAM Journal on Computing", vol. 36, n^{o} 6, 2007, p. 1777–1806.

[6] A. BOSTAN, F. MORAIN, B. SALVY, É. SCHOST. *Fast algorithms for computing isogenies between elliptic curves*, in "Mathematics of Computation", To appear, 2008.

[7] F. CHYZAK, M. DRMOTA, T. KLAUSNER, G. KOK. *The distribution of patterns in random trees*, in "Combinatorics, Probability & Computing", Published online. In press, 2007.

[8] P. DUMAS, H. LIPMAA, J. WALLÉN. *Asymptotic behaviour of a non-commutative rational series with a nonnegative linear representation*, in "Discrete Mathematics & Theoretical Computer Science", vol. 9, n^{o} 1, 2007, p. 247–274, http://www.dmtcs.org/dmtcs-ojs/index.php/dmtcs/article/view/478/1836.

[9] P. FLAJOLET, L. VEPSTAS. *On Differences of Zeta Values*, in "Journal of Computational and Applied Mathematics", To appear, 2008, http://arxiv.org/abs/math/0611332.

[10] É. FUSY. *Transversal structures on triangulations: combinatorial study and straight-line drawing*, in "Discrete Mathematics", To appear, 2008, http://arxiv.org/abs/math/0602163.

[11] E. FUSY, D. POULALHON, G. SCHAEFFER. *Bijective counting of plane bipolar orientations*, in "Electronic Notes in Discrete Mathematics", vol. 29, 2007, p. 283–287, http://www.sciencedirect.com/science/article/B75GV-4PCHPN5-1R/2/84e28461fe6672d2c0b27589b5b04f9e.

[12] É. FUSY, D. POULALHON, G. SCHAEFFER. *Dissections, orientations, and trees, with applications to optimal mesh encoding and to random sampling*, in "Transactions on Algorithms", To appear, vol. 4, n^{o} 2, April 2008.

[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", vol. 7, n^{o} 1, February 2007, p. 1–58.

[14] C. GOMEZ, B. SALVY. *Les Techniques de l'Ingénieur*, 34 pages, to appear, vol. Dossier AF1460, 2007.

[15] F. MEUNIER. *A combinatorial proof of a theorem of Freund*, in "Journal of Combinatorial Theory, Series A", To appear, 2008, 9 pages.

[16] F. MEUNIER, A. SEBÖ. *Parcours et coupes*, in "Graphes et applications, tome 2", Hermès Science Publications, 2007, p. 1–31.

Publications in Conferences and Workshops

[17] M. BODIRSKY, É. FUSY, M. KANG, S. VIGERSKE. *An unbiased pointing operator for unlabeled structures, with applications to counting and sampling*, in "ACM-SIAM Symposium on Discrete Algorithms (SODA)", 2007, p. 356–365.

[18] A. BOSTAN, F. CHYZAK, G. LECERF, B. SALVY, É. SCHOST. *Differential equations for algebraic functions*, in "ISSAC'07: Proceedings of the 2007 international symposium on Symbolic and algebraic computation", C. W. BROWN (editor), ACM Press, 2007, p. 25–32.

[19] A. BOSTAN, F. CHYZAK, F. OLLIVIER, B. SALVY, É. SCHOST, A. SEDOGLAVIC. *Fast computation of power series solutions of systems of differential equations*, in "ACM-SIAM Symposium on Discrete Algorithms (SODA)", Society for Industrial and Applied Mathematics, January 2007, p. 1012–1021.

[20] A. BOSTAN, C.-P. JEANNEROD, É. SCHOST. *Solving Toeplitz- and Vandermonde-like Linear Systems with Large Displacement Rank*, in "ISSAC'07: Proceedings of the 2007 international symposium on Symbolic and algebraic computation", C. W. BROWN (editor), ACM Press, 2007, p. 33–40.

[21] F. CHYZAK, A. QUADRAT, D. ROBERTZ. *OreModules: A symbolic package for the study of multidimensional linear systems*, in "Applications of Time Delay Systems", J. CHIASSON, J. J. LOISEAU (editors), Lecture Notes in Control and Information Sciences, vol. 352, Springer Berlin / Heidelberg, 2007, p. 233–264.

[22] P. FLAJOLET. *Analytic Combinatorics—A Calculus of Discrete Structures*, in "ACM-SIAM Symposium on Discrete Algorithms (SODA)", SIAM Press, 2007, p. 137–148.

[23] P. FLAJOLET, ÉRIC. FUSY, O. GANDOUET, FRÉDÉRIC. MEUNIER. *HyperLoglog: the analysis of a near-optimal cardinality estimation algorithm*, in "Analysis of Algorithms 2007 (AofA07)", P. JACQUET (editor), Discrete Mathematics and Theoretical Computer Science Proceedings, In press, 2007.

[24] P. FLAJOLET, É. FUSY, C. PIVOTEAU. *Boltzmann Sampling of Unlabelled Structures*, in "Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithmics and Combinatorics", D. APPELGATE (editor), Proceedings of the New Orleans Conference, SIAM Press, 2007, p. 201–211.

[25] É. FUSY. *Straight-line drawing of quadrangulations*, in "Graph Drawing 2006", Lecture Notes in Computer Science, Karlsruhe, Germany, September 18–20, 2006, vol. 4372, Springer-Verlag, 2007, p. 234–239.

[26] É. Fusy, F. Giroire. *Estimating the number of active flows in a data stream over a sliding window*, in "Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithmics and Combinatorics", D. Appelgate (editor), Proceedings of the New Orleans Conference, SIAM Press, 2007, p. 223–231.

Miscellaneous

[27] O. Bodini, É. Fusy, C. Pivoteau. *Random sampling of plane partitions*, Dec. 2007, http://arxiv.org/abs/0712.0111.

[28] J. M. Borwein, B. Salvy. *A proof of a recursion for Bessel moments*, June 2007, http://arxiv.org/abs/0706.1409, Submitted.

[29] É. Fusy. *A linear approximate-size random sampler for labeled planar graphs*, May 2007, http://arxiv.org/abs/0705.1287.

[30] É. Fusy, M. Kang, B. Shoilekova. *A Complete Grammar for Decomposing a Family of Graphs into 3-connected Components*, Nov. 2007, Submitted.

[31] A. Ivashenko, G. Boldina, A. Turmagambetova, M. Régnier. *Using profiles based on hydropathy properties to define essential regions for splicing*, December 2007, Submitted.

[32] M. Mezzarobba. *Génération automatique de procédures numériques pour les fonctions D-finies*, version 1.2, 89 pages, Rapport de stage, Master parisien de recherche en informatique, octobre 2007, http://www.eleves.ens.fr/home/mezzarob/info/m2/Mezzarobba_MScThesisMPRI2007-1.2.pdf.