

Projet Algo

Algorithmes

Rocquencourt

THÈME 2B

R *apport*
d'Activité

2002

Table des matières

1. Composition de l'équipe	1
2. Présentation et objectifs généraux	1
2.1.1. Thématique de recherche	2
3. Fondements scientifiques	3
3.1. Analyse d'algorithmes	3
3.2. Calcul formel	3
3.2.1. Structures combinatoires	4
3.2.2. Suites et fonctions spéciales	5
3.2.3. Séries et échelles asymptotiques	5
3.3. Algorithmique des séquences	5
4. Domaines d'application	6
4.1. Panorama	6
5. Logiciels	7
6. Résultats nouveaux	8
6.1. Analyse d'algorithmes	8
6.1.1. Combinatoire analytique	8
6.1.2. Analyse et optimisation d'algorithmes	9
6.2. Calcul formel	10
6.2.1. Combinatoire et asymptotique	10
6.2.2. Fonctions spéciales	10
6.2.3. Élimination	11
6.3. Algorithmique des séquences	11
7. Contrats industriels	13
7.1. Calcul formel	13
8. Actions régionales, nationales et internationales	14
8.1. Actions nationales	14
8.2. Actions financées par la Commission Européenne	14
8.3. Relations bilatérales internationales	14
8.4. Accueils de chercheurs étrangers	15
9. Diffusion des résultats	15
9.1. Animation de la communauté scientifique	15
9.2. Enseignement universitaire	16
9.3. Participation à des colloques, séminaires, invitations	16
10. Bibliographie	17

1. Composition de l'équipe

Responsable scientifique

Bruno Salvy [DR]

Responsable permanent

Philippe Flajolet [DR]

Assistante de projet

Virginie Collette [TR]

Personnel Inria

Frédéric Chyzak [CR]

Mireille Régnier [DR]

Ingénieur associé

Édouard Dolley [à partir du 1er octobre]

Collaborateurs extérieurs

Cyril Banderier [Bourse Post-doc INRIA, Sarrebruck, jusqu'au 30 septembre]

Philippe Dumas [professeur, Cl. Prépa. lycée Jean-Baptiste Say]

Pierre Nicodème [CR CNRS, Université d'Évry]

Maryse Pelletier [professeur, Université Paris VI]

Michèle Soria [professeur, Université Paris VI]

Brigitte Vallée [DR CNRS, Université de Caen]

Chercheurs Invités

Marni Mishna [Université du Québec à Montréal, du 10 septembre au 15 novembre]

Thomas Klausner [Technische Universität, Vienne, Autriche, du 17 novembre au 11 décembre]

Chercheur post-doctorant

Alexandre Sedoglavic [jusqu'au 30 septembre]

Doctorants

Alin Bostan [École polytechnique, co-encadrement]

Marianne Durand [ENS Paris]

Julien Fayolle [Université de Paris VI, à partir du 1er octobre]

Ludovic Meunier [École polytechnique]

Vincent Puyhaubert [ENS Cachan]

Mathias Vandenberg [Université de Bordeaux]

2. Présentation et objectifs généraux

L'objectif global du projet ALGO est l'analyse et l'optimisation fines de systèmes complexes discrets présentant une forte composante aléatoire. De nombreux problèmes de grande taille rentrent dans ce cadre tels l'étude quantitative d'algorithmes probabilistes sur des structures discrètes, ou l'étude des déviations par rapport à l'alea dans des séquences biologiques. La réalisation de cet objectif passe par la compréhension en profondeur de l'alea discret et des problèmes de nature mathématique que pose sa quantification. Cela suppose de dégager des méthodes générales pour obtenir des résultats présentés de manière exacte ou asymptotique. Ces résultats fournissent alors des informations très précises sur le comportement qualitatif ou quantitatif des systèmes étudiés.

Étant donné le caractère très systématique de l'approche poursuivie, des méthodes de décision réalisables en calcul formel font aussi partie des objectifs du projet. Cette approche est un moteur puissant de renouvellement qui conduit à la révision d'approches classiques dans le domaine des fonctions spéciales et des développements en séries. L'objectif est de disposer d'une algorithmique fiable et complète pour de grandes classes de problèmes précisément caractérisés ; voir déjà les bibliothèques GFUN et MGFUN assez largement utilisées

dans la communauté combinatoire et présentes dans les dernières versions de Maple. Les résultats sont susceptibles de nombreuses applications bien au delà du domaine de la modélisation combinatoire : ainsi une meilleure intégration des fonctions spéciales au calcul formel est visée, ce qui s'applique à de larges classes de problèmes dans les sciences de l'ingénieur.

2.1.1. *Thématique de recherche*

Le projet ALGO se donne comme objectif l'analyse en profondeur de l'aléa combinatoire et la recherche de ses lois générales. Ce thème est voisin par ses objectifs, mais dual par ses méthodes, de la modélisation des systèmes informatiques, laquelle repose de manière prédominante sur des mathématiques *a priori* continues. Ici, nous sommes dans le domaine des mathématiques discrètes. La combinatoire étant par définition l'étude des objets finis discrets, nous visons à développer une approche globale que l'on pourrait qualifier de « combinatoire statistique » par analogie avec la physique statistique. Le but est de décrire le comportement macroscopique visible de l'objet étudié, comme par exemple, le temps d'exécution sur un grand « ensemble » d'un certain algorithme, ou encore l'évolution d'un grand réseau vers ses régimes fondamentaux.

Il s'agit ainsi de caractériser les propriétés attendues (en moyenne, en probabilité ou en distribution) d'objets obéissant à des règles de combinaison finies, mais constituant de très grands ensembles. Ces situations se rencontrent sans cesse en informatique, un tri de seulement 100 éléments met en jeu 10^{158} configurations possibles, qui obéissent, avec une écrasante probabilité, à des règles fort précises. Les problèmes d'aléa combinatoire interviennent de manière essentielle en algorithmique. La conception de la plupart des algorithmes efficaces se fonde naturellement sur les cas attendus, en moyenne ou en probabilité, plutôt que sur une analyse pessimiste qui doit être réservée à des contraintes de type « temps réel ».

Le rôle de l'analyse de l'aléa combinatoire est renforcé par l'importance croissante des algorithmes dits « randomisés » (bien formalisés par Karp et Rabin depuis une vingtaine d'années) où il s'avère payant d'introduire volontairement le hasard dans le calcul. Ainsi, les tables de hachage constituent un substitut souvent très efficace aux arbres de recherche, les signatures accélèrent considérablement la recherche textuelle, les « skip lists » remplacent graduellement les arbres équilibrés dans de nombreuses applications. Parmi d'autres applications célèbres de cette « aléatorisation », on peut citer la construction de cryptosystèmes à clefs publiques qui utilise de manière très sûre des tests probabilistes ; ou encore la conception de protocoles de communication, faisant suite à Ethernet, mais avec une utilisation mieux contrôlée de l'aléa qui permettent de stabiliser et d'acheminer des trafics plus importants dans les réseaux. La problématique de l'algorithmique randomisée est poursuivie par exemple avec succès par le projet PRISME de Sophia-Antipolis dans le domaine de l'algorithmique géométrique et par le projet HIPERCOM de Rocquencourt dans le domaine des télécommunications.

Le développement récent du *calcul formel* a éclairé tout l'intérêt de cette approche généraliste. Il est apparu, en effet, vers le tournant des années 1990 qu'il était possible de décider mathématiquement de nombreuses propriétés de l'aléa combinatoire, ce par un calcul de nature essentiellement formelle. Au sein du projet, cela a donné lieu aux thèses de B. Salvy et P. Zimmermann (en 1991). Le programme de recherche correspondant à ces aspects est loin d'être achevé. De nombreux problèmes de portée générale en calcul formel ont été mis en évidence. Citons principalement ici, comme directions nouvelles, l'asymptotique automatique (B. Salvy), les méthodes mixtes symboliques numériques (B. Salvy), et la preuve automatique d'identités combinatoires (F. Chyzak, B. Salvy).

C'est dans ce contexte qu'ont pu être résolues au fil des ans diverses conjectures correspondant à des analyses précises d'algorithmes tels que le dimensionnement en hachage dynamique (M. Régnier, P. Flajolet), la performance des arbres quadrants pour la recherche multidimensionnelle (P. Flajolet, B. Salvy), ou le comportement probabiliste des meilleures méthodes de recherche de motifs (M. Régnier). D'autres applications typiques sont constituées par les algorithmes d'estimation probabiliste en bases de données, l'allocation mémoire partagée (P. Flajolet) et les protocoles en arbre (P. Flajolet et P. Jacquet). Les travaux du groupe sont ainsi assez largement repris dans la littérature scientifique spécialisée (livres ou articles). Ils nous valent de nombreuses coopérations internationales avec des universités comme celles de Barcelone, Princeton, Purdue, Stanford, Vienne, Washington, Waterloo, etc.

3. Fondements scientifiques

3.1. Analyse d'algorithmes

Participants : Cyril Banderier, Marianne Durand, Philippe Flajolet, Mireille Régnier, Bruno Salvy, Michèle Soria, Brigitte Vallée.

Mots clés : *analyse d'algorithmes, analyse combinatoire, combinatoire analytique, structures aléatoires discrètes, arbre d'index, algorithme de hachage, algorithme géométrique, loi limite.*

L'analyse d'algorithmes vise à une quantification précise des principaux algorithmes et structures de données de l'informatique. Elle repose sur une étude approfondie de l'aléa discret. Une approche originale développée au sein du projet se fonde sur la « combinatoire analytique ». Les applications en sont les principaux problèmes de recherche de l'information dans de grands volumes de données et l'algorithmique de la communication.

L'analyse d'algorithmes ou de structures de données dépend très étroitement d'une évaluation et d'une quantification précise de l'aléa discret. Il s'agit en effet de caractériser le comportement des principaux paramètres de structures combinatoires d'arbres, de mots, de graphes ou d'allocations aléatoires. Dans ce domaine, les travaux du projet ALGO s'appuient sur une approche originale fondée sur :

- les méthodes symboliques en analyse combinatoire qui permettent de disposer de méthodes systématiques de mise en équation par séries génératrices ;
- les méthodes asymptotiques complexes fondées notamment sur l'analyse de singularité.

L'application de ces idées aux séquences textuelles ou génétiques fait l'objet de la section 3.3. L'automatisation de ces méthodes visant à l'analyse complète de modèles complexes est traitée à la section 3.2.

Traditionnellement, l'analyse d'algorithmes propose des prédictions quantitatives de comportement moyen, c'est-à-dire des analyses d'espérances de coûts. Une évolution importante qui commence au début des années 1990 avec la thèse de doctorat ès-sciences de M. Soria consiste à aborder des études en distribution. Ceci permet non seulement de connaître les comportements moyens attendus mais encore de prédire les profils de coûts sous la forme de distributions limites ou de bornes de grandes déviations. En d'autres termes, on parvient depuis peu à quantifier très précisément les compromis de type risque-efficacité. Du point de vue mathématique, les travaux établissent un pont intéressant entre approches probabilistes classiques et approches issues de la combinatoire analytique. Une monographie intitulée « Analytic Combinatorics » par P. Flajolet et R. Sedgewick est en préparation (500 pages sont déjà rédigées et publiées en rapports de recherche INRIA). Cette monographie fait partie d'un programme de recherche visant à établir un corpus cohérent de méthodes pour le domaine de la combinatoire analytique et pour la quantification de l'aléa discret. Les méthodes qui sous-tendent ce programme ont de nombreuses applications en analyse d'algorithmes, parmi lesquelles sont traités spécifiquement dans le projet : les structures de données fondamentales comme les tables de hachage, les arbres de recherche uni- ou multi-dimensionnels ; l'algorithmique de la recherche textuelle et du traitement de séquences génomiques (voir la section 6.3) ; les arbres digitaux liés à des problèmes de nature très générale en théorie analytique de l'information (voir la section 6.1). Ces méthodes inspirent également une partie des recherches en calcul formel (section 6.2) par les problèmes que pose l'automatisation de la manipulation de grands modèles combinatoires.

3.2. Calcul formel

Participants : Frédéric Chyzak, Ludovic Meunier, Marni Mishna, Bruno Salvy, Alexandre Sedoglavic.

Mots clés : *échelles asymptotiques, génération aléatoire, fonctions spéciales, élimination polynomiale, bases de Gröbner.*

Le projet ALGO développe des algorithmes de calcul formel qui permettent : le traitement de modèles combinatoires récurrents ; l'analyse asymptotique automatique de nombreuses classes de problèmes ; le traitement

de fonctions spéciales et de suites combinatoires de manière unifiée, et notamment leur intégration ou leur sommation. Ces algorithmes sont dans la plupart des cas validés par des implantations. Leur développement est motivé par l'automatisation d'une approche combinatoire à l'analyse d'algorithmes.

Les trois étapes fondamentales de l'analyse d'algorithmes telle qu'elle est pratiquée au projet ALGO sont la modélisation combinatoire, la manipulation de séries génératrices et l'analyse asymptotique. Chacune de ces étapes requiert des capacités de calcul symbolique importantes, tant pour l'application des méthodes symboliques que pour l'expérimentation. Ce besoin explique l'importance de l'activité en calcul formel au sein du projet. L'objectif à long terme est de systématiser et d'automatiser ces trois étapes.

Au cours des années ont été élaborés de nombreux algorithmes et programmes. Les domaines mathématiques permettant l'expression et la preuve de nos algorithmes sont la combinatoire, l'analyse asymptotique, l'algèbre différentielle (corps de Hardy et anneaux de polynômes non commutatifs). Notre travail en calcul formel est ainsi complémentaire de celui des projets GALAAD, SPACES et CAFE consacrés aux systèmes polynomiaux et aux équations différentielles.

Comme pour l'analyse d'algorithmes, l'approche du projet aux algorithmes du calcul formel est globalisante et unificatrice. La résolution de problèmes appliqués mettant en œuvre de la combinatoire ou de l'analyse d'algorithmes est abordée à un niveau de généralité qui permet le développement d'une algorithmique de portée large. Ainsi, les travaux sur la combinatoire fournissent des générateurs aléatoires efficaces susceptibles de nombreuses applications ; les travaux sur l'analyse d'algorithmes ont abouti au développement d'une algorithmique d'échelles asymptotiques très générale, dont le besoin s'était fait sentir en intégration numérique et en physique mathématique. Les outils développés dans le projet touchent maintenant un public assez large d'utilisateurs du calcul formel intéressés tant par la combinatoire que par les fonctions spéciales, les manipulations de séries ou l'analyse asymptotique. La diversité de ce public est encore accrue par la présence de certains de nos programmes dans les bibliothèques du système de calcul formel MAPLE, l'un des deux principaux systèmes généralistes actuellement disponibles.

On peut distinguer trois grandes directions de travail dans notre activité en calcul formel : les structures combinatoires, les suites et fonctions spéciales et l'asymptotique automatique.

3.2.1. Structures combinatoires

Un langage de description généralisant les grammaires *context-free* permet d'exprimer des objets aussi divers que permutations, arbres binaires, arbres généraux, partitions d'entiers ou d'ensembles, graphes fonctionnels ou molécules chimiques, par exemple carbures ou alcools. À partir d'une description de structure décomposable, il est possible de (i) compter efficacement le nombre d'objets d'une certaine taille répondant à la spécification ; (ii) produire des fonctions de génération aléatoire uniforme de faible complexité - utiles pour des tests statistiques ; (iii) produire des fonctions de génération exhaustive de ces objets - utiles pour des tests de robustesse de procédures ; (iv) produire des itérateurs, c'est-à-dire des fonctions permettant d'accéder successivement à tous les objets d'une certaine taille, mais sans les stocker tous en mémoire ; (v) calculer des équations satisfaites par les séries génératrices d'énumération de ces objets - utiles pour la phase d'analyse asymptotique.

Le programme $\Lambda\Upsilon\Omega$ réalisé au début des années 90 par B. Salvy et P. Zimmermann fournissait une partie de ces fonctionnalités, mais sa portabilité et ses fonctionnalités étaient limitées par l'usage de CAML en conjonction avec MAPLE. L'objectif poursuivi depuis plusieurs années est de tirer parti de l'expérience acquise avec $\Lambda\Upsilon\Omega$ pour réaliser une version (COMBSTRUCT) entièrement intégrée en MAPLE, en mettant l'accent sur la modularité, la robustesse et la souplesse d'emploi.

Une vitrine en ligne de certaines de ces fonctionnalités est par ailleurs accessible sur le web sous le nom d'*Encyclopedia of combinatorial structures*. Cette encyclopédie regroupe environ un millier de structures combinatoires pour lesquelles sont données spécification, énumération, et lorsque c'est possible asymptotique, série génératrice, récurrence et forme exacte des coefficients. Cette encyclopédie est par ailleurs liée à l'*Encyclopedia of integer sequences* maintenue par N. Sloane (Bell Labs).

3.2.2. Suites et fonctions spéciales

Selon l'origine combinatoire du problème, les séries génératrices que l'on est amené à étudier peuvent être données sous des formes diverses. Elles peuvent être connues sous forme explicite mais elles sont le plus souvent décrites sous forme implicite par une ou plusieurs équations, fonctionnelles, différentielles ou aux différences. De même, leurs coefficients peuvent vérifier des récurrences de natures diverses. Manipuler ces fonctions définies implicitement nécessite des innovations théoriques, ainsi qu'un important effort d'implantation. Ce thème de recherche touche aux fondements du calcul symbolique, où il apparaît qu'il est paradoxalement souvent plus facile de traiter une fonction lorsqu'elle est représentée comme solution d'équations que lorsqu'elle est représentée sous forme close. En particulier, les questions de simplification et de formes normales qui sont une des difficultés majeures rencontrées par l'utilisateur trouvent une bien meilleure réponse dans ce contexte.

Le cas des solutions d'équations différentielles ou de récurrences *linéaires* attire beaucoup l'attention de la communauté de combinatoire et de calcul formel. De nombreuses suites et fonctions spéciales sont définies par de telles équations, qui bénéficient d'une algorithmique très riche, implantée en grande partie dans le *package* GFUN développé par B. Salvy et P. Zimmermann (projet SPACES). Une *encyclopedia of special functions* reposant sur cette algorithmique est actuellement développée par L. Meunier.

Le pendant multivarié de ces travaux est extrêmement fructueux, puisqu'il permet d'envisager une approche algorithmique de nombreux calculs avec des polynômes orthogonaux, des fonctions spéciales, ou plus généralement des fonctions, suites, séries ou distributions définies par un *système* d'équations linéaires à coefficients polynomiaux. Le travail mené par F. Chyzak considère des systèmes d'opérateurs linéaires qui peuvent être différentiels, aux différences, aux q -différences, ou de nombreux autres types. Il s'avère que les opérations d'addition, de produit, d'intégration ou de sommation peuvent s'effectuer algorithmiquement par le biais de techniques d'élimination sophistiquées dans des algèbres non commutatives d'opérateurs linéaires. Le *package* MGFUN développé par F. Chyzak valide et motive les recherches théoriques correspondantes.

3.2.3. Séries et échelles asymptotiques

Les besoins de la combinatoire analytique en matière de développements asymptotiques dépassent les capacités actuelles des systèmes de calcul formel. En effet, les calculs de coûts moyens et plus encore de variance donnent systématiquement lieu à des annulations non seulement dans les premiers termes des développements mais aussi dans l'ordre de grandeur exponentiel des croissances. La construction automatique des échelles asymptotiques nécessaires et le calcul avec ces échelles pose de nombreux problèmes sur lesquels le calcul formel est en progrès rapide. Les premiers travaux sur ce sujet datent des années 90. En 1988, G. Gonnet et K. Geddes (créateurs du système MAPLE) proposent un modèle permettant de traiter des formules de complexité proche de la formule de Stirling. Puis en 1990, J. Shackell (University of Canterbury) publie un algorithme qui permet de déterminer *de manière garantie* la limite des fonctions exp-log (fonctions de base de l'asymptotique). L'année suivante, la thèse de B. Salvy propose une première implantation de développements asymptotiques dans des échelles asymptotiques générales. Depuis lors, grâce à une série de travaux en commun de B. Salvy et J. Shackell, des classes de plus en plus vastes de fonctions ont reçu un algorithme permettant le calcul asymptotique de manière garantie. Ces progrès n'ont pas toujours été suivis d'implantation, mais une nouvelle structure de données pour les développements asymptotiques, les multiséries, a vu le jour récemment. L'algorithmisation de cette structure de données conduit à revisiter et prolonger les algorithmes anciens, et mène à la réalisation de prototypes permettant déjà de résoudre des problèmes du niveau de la recherche. Cette structure de données a été intégrée cette année dans le système MAPLE. Là encore, on s'attaque à une brique essentielle des systèmes de calcul formel, puisque les séries sont la structure de données efficace utilisée pour manipuler les polynômes, eux-mêmes à la base de nombreuses structures de données.

3.3. Algorithmique des séquences

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

Mots clés : combinatoire des mots, séquences, recherche de motifs, génome.

L'objet des recherches sur les séquences est la conception de nouveaux algorithmes, l'obtention de la complexité moyenne de ces algorithmes, l'implantation et l'application à l'algorithmique de certains résultats statistiques. Plus généralement, nous développons une théorie analytique de l'information qui s'appuie sur la combinatoire, les probabilités et l'analyse. L'algorithmique des séquences ou objets textuels couvre des domaines d'application variés (compression des données textuelles et séquences biologiques). Ce sujet comprend d'abord des recherches algorithmiques. Il s'agit de trouver efficacement un motif ou un ensemble de motifs dans un texte. Cet ensemble peut en particulier être l'ensemble des mots voisins d'un mot donné, à un nombre d'erreurs près, où le type et le nombre d'erreurs autorisées sont déterminés par l'application. Une nouvelle classe d'algorithmes dits d'extraction de motifs, émerge dans les recherches sur le génome. On extrait des motifs à signification biologique particulière, non connus à l'avance. Ce sont par exemple des mots à fréquence exceptionnelle : sur-représentés ou sous-représentés pour un modèle probabiliste donné. On peut aussi rechercher une expression régulière caractéristique d'une famille donnée. Ces algorithmes testent sur une même séquence un grand nombre de candidats potentiels et la rapidité du calcul statistique, et sa précision, apparaissent essentielles.

D'un point de vue méthodologique, des théorèmes probabilistes trouvent des applications naturelles dans l'étude des séquences. Plus précisément, nous avons mis en évidence différents types de processus de renouvellement, la loi limite étant généralement gaussienne ; le calcul effectif des paramètres de coût peut être très délicat et les outils combinatoires et analytiques permettent pour cette classe de problèmes les calculs effectifs des distributions. Les formules obtenues sont aussi valables dans le domaine fini. Enfin, l'étude des queues de distribution précise l'évaluation de la sur-représentation d'un motif. L'analyse asymptotique fournit une expression exacte et calculable de la fonction de taux, au sens de la théorie des grandes déviations. La précision de calcul ainsi atteinte est meilleure que celle d'un calcul exact, numériquement instable. Parallèlement, nous développons des outils de calcul dans le modèle probabiliste markovien. La complexité des évaluations de performances dans le cas markovien provient du nombre de cas différents à considérer. Nous définissons pour chaque problème des langages caractéristiques dont la contribution au coût total de l'algorithme est calculable. Ceci équivaut à une agrégation des états de l'automate associé. Les résultats s'appliquent à la compression et à la recherche de motifs avec erreurs ou de motifs exceptionnels dans les textes (DosDNA, reconnaissance de gènes ou de signaux de régulation, etc.). Ils permettent aussi d'établir les domaines d'efficacité des différents algorithmes de recherche de similarités dans des bases de données protéiques.

4. Domaines d'application

4.1. Panorama

Les applications de nos travaux sur les structures combinatoires sont la modélisation et l'étude des systèmes discrets complexes et des réseaux de télécommunication. Les applications visées par l'analyse d'algorithmes sont les méthodes d'accès rapide à des informations structurées, une algorithmique rapide du calcul formel et le traitement statistique des séquences biologiques.

Nos domaines de recherche en calcul formel sont : les structures combinatoires, les suites et fonctions spéciales et l'analyse asymptotique. Les applications de nos travaux sur les structures combinatoires sont la modélisation et l'étude de systèmes discrets complexes. Nos résultats sur les suites et fonctions spéciales débouchent sur la manipulation par le calcul formel de fonctions spéciales intervenant de manière classique en physique mathématique et en mécanique. Nos travaux sur l'asymptotique devraient permettre à long terme de faire le pont entre le calcul numérique et le calcul formel : le calcul numérique, robuste en l'absence de singularités, pourrait être complété par une étude formelle fine au voisinage des singularités débouchant sur de la production de code numérique robuste dans ces zones sensibles.

5. Logiciels

La bibliothèque COMBSTRUCT a été conçue et développée par le projet ALGO (P. Flajolet, M. Mishna, B. Salvy, E. Murray) en liaison avec le projet SPACES de Nancy (P. Zimmermann). Elle fait l'objet d'une collaboration régulière avec les groupes de Waterloo (Université et Compagnie WMI) et une version assez récente est intégrée au système Maple. Elle permet actuellement la génération aléatoire ou exhaustive, le calcul automatique de dénombrements et de séries génératrices, et est à ce titre une aide de portée générale pour la simulation et le test systématique de modèles combinatoires. Disons qu'en l'état actuel, et sur son créneau, son expertise est de l'ordre de celle d'un étudiant en début de 3ème cycle. La dernière version de COMBSTRUCT comporte une importante extension à une classe de grammaires attribuées, développée par M. Mishna, qui permet de calculer automatiquement les séries génératrices de coût d'une large classe d'algorithmes. La version actuelle représente environ 8 500 lignes de code (300 ko).

Le *package* GFUN développé par B. Salvy et P. Zimmermann (projet SPACES) fournit de nombreux outils de manipulations de suites et de fonctions, à commencer par une fonction de traduction qui part de la forme close d'une fonction et produit une équation différentielle linéaire dont cette fonction est solution (lorsqu'une telle équation existe). Cette fonctionnalité, qui effectue précisément le chemin inverse de celui vers lequel se précipitent nombre d'utilisateurs, permet ensuite de calculer des développements en série de manière plus rapide qu'avec la forme close ; elle permet également la localisation des singularités et le calcul des comportements au voisinage des singularités. Le *package* GFUN comporte actuellement environ 3 300 lignes de code Maple. Une version récente est intégrée à la distribution Maple. Il a fait l'objet d'une revue très positive dans *Computing Reviews* et est incorporé au *superseeker* de N. Sloane aux Bell Laboratories, accessible sur le Web, qui détermine de nombreuses suites d'après leurs premiers termes. Une extension de GFUN pour produire automatiquement des procédures efficaces d'évaluation numérique en précision arbitraire fera partie de la prochaine version.

L'*Encyclopedia of combinatorical structures* est un *package* Maple qui repose sur la combinaison de COMBSTRUCT, GFUN et GDEV et est doté d'un mode d'interrogation via le Web à l'adresse <http://algo.inria.fr/encyclopedia>.

Dans le même esprit, L. Meunier a développé une encyclopédie en ligne de fonctions spéciales univariées, prolongement naturel de GFUN, où toutes les informations sur les fonctions spéciales (développement en série et asymptotique y compris l'expression du terme général lorsque c'est possible, procédure d'évaluation numérique, développement sur des familles de polynômes orthogonaux, graphiques, transformées intégrales) sont produites automatiquement à partir d'une équation différentielle et de ses conditions initiales.

Les algorithmes développés par F. Chyzak sont implantés dans une bibliothèque Maple du nom de MGFUN partiellement intégrée dans la distribution grand public de Maple. Depuis quelques versions, les utilisateurs de Maple disposent ainsi d'une bibliothèque pour la manipulation d'opérateurs linéaires, ainsi que d'une nouvelle bibliothèque pour calculer des bases de Gröbner, capable de traiter aussi bien des polynômes que des opérateurs non commutatifs. C'est désormais cette dernière qui est employée pour réaliser l'élimination polynomiale dans tout le logiciel Maple. Par ailleurs, la simple présence dans Maple de routines de calcul sur les opérateurs linéaires a ouvert la voie à l'implantation, jusqu'alors impossible, de toute une nouvelle génération d'algorithmes pour la sommation et l'intégration symboliques, et plus généralement d'algorithmes récents pour la manipulation de représentations implicites de suites et fonctions spéciales. Suite à l'intégration de MGFUN, la société WMI qui développe Maple a accentué son effort dans cette direction. L'ensemble de la réalisation logicielle correspondant à la bibliothèque MGFUN est constitué de 14 000 lignes de code (520 Ko), accompagnées d'une quantité équivalente de documentation et de jeux de tests ; elle apporte à l'utilisateur, plus d'une cinquantaine de nouvelles fonctions. Afin de rendre l'utilisation de MGFUN plus transparente aux utilisateurs dont les problèmes ne requièrent pas toute la puissance du *package*, une extension de MGFUN a été réalisée en collaboration avec C. Germa (ancien stagiaire dans le projet). Cette extension détermine automatiquement l'algèbre et les opérateurs avec lesquels effectuer les calculs de somme ou d'intégrale demandés par l'utilisateur. Le *package* MGFUN a par ailleurs été choisi pour ses primitives de calcul de bases

de Gröbner pour le *package* DESING développé au RISC (Université de Linz, Autriche) et disponible sur le Web (<http://www.risc.uni-linz.ac.at/projects/basic/adjoints/blowup/>).

L'ensemble des *packages* ci-dessus est intégré en une unique bibliothèque du nom d'ALGOLIB, et diffusée par les pages web du projet ALGO. L'installation par les utilisateurs en est ainsi grandement simplifiée, tout en permettant une meilleure interaction entre les *packages*. Plus de 150 utilisateurs se sont inscrits volontairement à une liste de diffusion pour recevoir les annonces de nouvelles versions d'ALGOLIB.

6. Résultats nouveaux

6.1. Analyse d'algorithmes

Participants : Cyril Banderier, Marianne Durand, Philippe Flajolet, Mireille Régnier, Bruno Salvy, Michèle Soria, Brigitte Vallée.

Les recherches en 2002 dans ce domaine se sont poursuivies sur deux fronts. D'une part, les méthodes générales de quantification de l'aléa discret, sont regroupées sous l'intitulé *Combinatoire analytique*. D'autre part, le paragraphe intitulé *Analyse et optimisation d'algorithmes* démontre tout l'intérêt d'une approche globale dans la résolution effective de nombreux modèles quantitatifs liés à diverses algorithmiques spécifiques ainsi qu'aux structures de données correspondantes. Les recherches du projet ALGO se placent dans ce cadre général, et la partie décrite ici constitue le « tronc commun » à de nombreux autres travaux conduits dans l'équipe. Ces études plus fondamentales sous-tendent ainsi les applications aux séquences textuelles ou génétiques (section 6.3) et à certains problèmes de modélisation de réseaux (voir le rapport d'activité du Projet RAP). Enfin, l'automatisation de méthodes visant à l'analyse complète de modèles complexes est intimement liée au calcul formel et traitée à la section 6.2.

6.1.1. Combinatoire analytique

Un premier lot de travaux conçus ou publiés en l'an 2002 a traité à l'approfondissement des méthodes fondamentales de la combinatoire analytique. Deux aspects sont couverts : l'aspect combinatoire qui repose sur une élaboration des méthodes de dénombrement exact, et l'aspect analytique, c'est-à-dire asymptotique, qui permet d'extraire des prédictions quantitatives à la fois précises et simples dans leurs formes.

Un premier travail de P. Flajolet et R. Sedgewick (Princeton) concrétisé par un rapport de 186 pages [20] offre une présentation unifiée de l'analyse combinatoire classique ainsi que des méthodes de dénombrement par séries génératrices. Ce rapport est la première brique d'un ouvrage de référence, en préparation, lequel s'intitule « *Analytic Combinatorics* ». Les modèles couverts incluent les modèles rationnels (utiles à l'analyse des mots, motifs, et séquences), les modèles algébriques (liés par exemple à diverses configurations géométriques), ainsi que de nombreux modèles de l'aléa dans les arbres, graphes, permutations, partitions, et allocations. La formalisation sert également de base théorique aux bibliothèques combinatoires issues du projet, et qui sont implantées dans le système de calcul formel MAPLE. Une application de ces méthodes à la multiconnectivité dans les réseaux aléatoires fait l'objet de l'article [7], développé en collaboration avec une équipe du CTI à Patras. L'article de Banderier (CNRS, Villetaneuse) et Flajolet [4] montre une autre application des méthodes combinatoires au dénombrement de chemins plans.

Dans une conférence invitée à l'*International Congress of Mathematicians* (ICM, Beijing, août 2002, cf [16]), P. Flajolet résume les principaux progrès accomplis au cours de la décennie écoulée dans le domaine de l'analyse de singularité. La démarche consiste à interpréter les séries génératrices de l'analyse combinatoire comme des fonctions analytiques (holomorphes) et examiner le reflet des constructions combinatoires sur la géométrie des singularités de ces fonctions. De nombreux résultats découlent de cette approche : forme asymptotiques des dénombrements, lois limites de paramètres, classification des principaux schémas en grandes classes de processus combinatoires partageant des propriétés communes. Le projet a un rôle de leader dans ces questions. Une application originale des méthodes d'analyse de singularité à un problème de théorie de l'information se trouve dans l'étude [8], laquelle a été menée avec W. Szpankowski (Purdue University), collaborateur régulier des projets ALGO et HIPERCOM.

Que se passe-t-il lorsque les fonctions génératrices cessent d'être analytiques ? L'article [19] montre que l'on peut encore exploiter les propriétés analytiques complexes sur les séries divergentes issues de l'analyse de la connectivité dans les graphes aléatoires. Procéder ainsi ouvre la voie de l'analyse dans l'étude des modèles classiques d'Erdős et Rényi. Au plan méthodologique, nous faisons appel à la méthode dite des cols coalescents, laquelle est issue des mathématiques appliquées. L'étude [19] met en évidence le rôle important des fonctions d'Airy dans les problèmes de transition de phase en combinatoire, ce qui est un des thèmes d'intérêt récurrents du projet. Ainsi la connectivité du graphe aléatoire, la dégénérescence des tables de hachage pleines, le parcours d'arbres, l'évolution des piles sous l'effet d'insertions et suppressions aléatoires, et la multiconnectivité des cartes aléatoires se retrouvent sous un cadre commun, celui des phénomènes de coalescence de cols.

À travers une apparente diversité technique, ces recherches procèdent d'une entreprise globale qui vise à délimiter précisément le champ des méthodes de la combinatoire analytique. Les objets traités sont des objets de base de l'informatique : arbres, chemins, allocations aléatoires, mots et quantité d'information, hachage, gestion de caches, graphes, configurations géométriques, etc.

6.1.2. Analyse et optimisation d'algorithmes

Les études de combinatoire analytique sont étroitement liées à la compréhension en profondeur des propriétés de l'aléa discret ; elles sont dans le même temps très largement motivées par les besoins d'analyse et d'optimisation des principaux algorithmes de base de l'informatique.

L'algorithme de « tri-rapide » (ou *Quicksort*) est le procédé de tri généraliste utilisé par excellence dans de nombreux systèmes informatiques, notamment à cause de ses boucles internes très rapides. Si l'analyse en moyenne de l'algorithme de base est déjà ancienne (D. Knuth 1973, R. Sedgewick 1975), on assiste actuellement à un renouveau lié à des propriétés dégagées au cours des dernières années, telles l'existence d'une loi limite (M. Régnier) ou la construction de bornes de grandes déviations qui quantifient les probabilités d'événements exceptionnels. M. Durand [6] s'est attachée à développer une analyse complète de la version de Quicksort qui est utilisée dans les systèmes UNIX FreeBSD. Il s'agit très exactement de la version finement optimisée par Bentley sur la base empirique de simulations, pour laquelle l'analyse apporte une validation théorique plaisante, à quelques pour cent près.

V. Puyhaubert explore l'impact possible des méthodes analytiques sur les instances de problèmes « difficiles », c'est-à-dire, NP-complets (partitionnement entier, satisfaisabilité de formules booléennes, par exemple). Des études empiriques issues de simulations ainsi que différents modèles heuristiques suggèrent l'existence de seuils nets séparant les régions facilement satisfaisables des régions clairement contradictoires. La région correspondant à une « transition de phase » semble bien être celle où se concentre toute la complexité du problème. Ces questions sont liées à la résolution de contraintes ; la quantification précise de ces phénomènes est alors d'un intérêt pratique en optimisation combinatoire évident puisqu'elle doit conduire à concevoir des filtres décidant efficacement des stratégies à utiliser au mieux dans chaque cas. V. Puyhaubert dispose déjà d'une approche combinatoire et analytique unifiée qui fournit de manière transparente les meilleures bornes supérieures connues au problème de la 3-satisfaisabilité (3-SAT).

Parmi les structures de données les plus fondamentales de l'informatique, l'arbre digital ou *trie* se situe en bonne place. Cette structure intervient dans l'accès rapide à des données textuelles (voir Section 6.3), en algorithmique probabiliste des bases de données, et même dans la conception d'une gamme de protocoles de communication plus efficaces et robustes qu'ETHERNET ; voir les rapports d'activité du projet HIPERCOM, des dernières années. Les arbres suffixes sont une structure très usitée, en particulier dans le domaine de la compression de données (comme dans le programme *gzip*). Dans la ligne de son mémoire [22], Julien Fayolle vise à obtenir une détermination asymptotique du comportement des paramètres d'un arbre suffixe dans le cas où les symboles constituant le texte sont émis par une source dynamique. Le concept récent de source dynamique a été introduit par B. Vallée, il généralise le cas où les symboles sont produits avec une probabilité initiale donnée et ne tiennent pas compte des précédents symboles. Des travaux voisins concernant les motifs dans les textes aléatoires [23][9] dus à P. Flajolet, P. Nicodème, W. Szpankowski (Purdue), B. Salvy et B. Vallée sont présentés dans la section « Algorithmiques des séquences ».

Marianne Durand a développé l'analyse d'une structure de donnée nommée `JumpList` créée par Hervé Bronnimann (New York) et Frédéric Cazals (Projet PRISME, Sophia-Antipolis). Cette structure de donnée très simple consiste à ajouter des pointeurs de saut appelés « *jumps* » à une liste simplement chaînée. L'analyse du coût d'accès d'un élément dans cette structure montre que ce coût est plus faible que le coût d'accès dans un arbre binaire de recherche ; la structure permet la gestion très efficace des recherches par intervalle et la gestion des tableaux dynamiques.

Enfin, la génération aléatoire est l'un des domaines anciens d'activité du projet (en liaison notamment avec P. Zimmermann du Projet SPACES). C'est aussi l'une des bases historiques de la coopération entre l'INRIA et la société canadienne WMI qui développe le système MAPLE, voir la bibliothèque COMBSTRUCT et la section 6.2, <http://algo.inria.fr/libraries/>. L'article [3] issu d'une collaboration multipartie entre l'INRIA, le Labri (Bordeaux), l'Université de Versailles et le LRI (Orsay), présente des résultats sur le tirage aléatoire d'objets combinatoires par la méthode des « arbres de génération ». La communication au colloque ICALP'02 [15] propose une approche radicalement nouvelle à la génération aléatoire d'objets structurés complexes qui donne lieu dans la plupart des cas à des algorithmes dont la complexité en nombre *total* d'opérations est tout simplement *linéaire* (au lieu de quadratique ou cubique, selon les algorithmes classiques du domaine). Il s'agit d'une nouvelle gamme de générateurs fondés sur des principes inspirés de la mécanique statistique (principe de Boltzmann-Gibbs) et reposant sur de simples évaluations numériques de précision modérée. L'année 2003 devrait voir un développement rapide de ces idées étant donné le caractère pratique extrêmement prometteur de ces modes radicalement nouveaux de génération d'ensembles de tests. Ces travaux sont menés en étroite collaboration avec G. Schaeffer du Projet ADAGE à l'INRIA-Lorraine, P. Duchon à Bordeaux, et G. Louchard à l'Université Libre de Bruxelles.

6.2. Calcul formel

Participants : Alin Bostan, Frédéric Chyzak, Thomas Klausner, Ludovic Meunier, Marni Mishna, Bruno Salvy.

6.2.1. Combinatoire et asymptotique

L'automatisation de la combinatoire analytique pour l'analyse automatique de la complexité d'algorithmes aborde désormais les questions de distributions limite des coûts. La première étape consiste à obtenir des représentations de séries génératrices multivariées, l'une des variables marquant la taille des objets et les autres les coûts de chacune des procédures à analyser. L'étape suivante consiste à extraire de ces représentations des informations quant au comportement singulier des séries génératrices, et surtout préciser la dépendance de ce comportement vis-à-vis des diverses variables. De nombreuses questions liées à une classification complète des comportements singuliers attendus sont encore en friche, mais des premiers pas vers l'obtention automatique du caractère gaussien de nombreux paramètres ont été effectués par T. Klausner. Le caractère effectif de ces résultats repose en grande partie sur les outils asymptotiques développés depuis de nombreuses années par B. Salvy, qui permettent désormais d'aborder des calculs que leur complexité rend difficilement amenable à un traitement manuel. Ces travaux ont été étendus cette année dans [21] aux développements asymptotiques à coefficients oscillants, dans une collaboration avec J. Shackell (U. Kent, Canterbury).

6.2.2. Fonctions spéciales

L. Meunier développe une *Encyclopedia of Special Functions*, produite *automatiquement* à l'aide d'algorithmes et d'implantations d'outils issus notamment des travaux menés par B. Salvy et P. Zimmerman (package GFUN) d'une part, et par F. Chyzak (package MGFUN) d'autre part. Cette encyclopédie rassemble des identités, des formules et des graphes qui sont calculés automatiquement à partir de la donnée d'une équation différentielle linéaire et de conditions initiales. Tout le processus de production étant automatisé, les étapes difficiles et coûteuses de vérification individuelle de chaque formule sont supprimées. Disponible sur le web (<http://algo.inria.fr/esf>), cette encyclopédie a ainsi un rôle de vitrine pour une partie des *packages* du projet.

Une collaboration de F. Chyzak avec P. Paule concerne la rédaction du chapitre sur les méthodes de calcul formel pour les fonctions spéciales, dans le cadre du projet *Digital Library of Mathematical Functions* (DLMF)

du *National Institute of Standards and Technology* (bureau des standards américain). Ce projet ambitieux vise à fournir une nouvelle édition du « *Handbook of Mathematical Functions* », formulaire de référence depuis 1962 et peut-être l'ouvrage déjà le plus cité dans l'histoire des publications scientifiques, qui sera à la fois disponible en version imprimée (environ 1 000 pages) et sous forme électronique (CD-rom et site Web, voir <http://dlmf.nist.gov/>). L'achèvement de ce projet est prévu fin 2002. Contrairement à notre *Encyclopedia of Special Functions*, ce projet est entièrement statique. Un but de plus long terme du NIST sera cependant de faire un usage complet de modes de communication avancés et de moyens de calcul automatisés, de façon à présenter non seulement des données statiques, mais aussi de l'information dynamique produite à la demande, telle que des graphes de fonctions, des tables de valeurs numériques et même, des tables d'identités et de transformations symboliques. Le caractère autoritaire du formulaire existant et son orientation vers les applications en sciences, statistiques, ingénierie et calculs seront préservés ; mais sa valeur utilitaire se trouvera largement étendue, dépassant les limitations traditionnelles des media imprimés, et faisant du DLMF un véhicule pour révolutionner la pratique et la diffusion des mathématiques appliquées en général. Des discussions ont déjà eu lieu entre des membres du projet ALGO et le NIST sur de possibles coopérations sur ces thèmes dans le futur.

Une ouverture récente des travaux de F. Chyzak et B. Salvy est l'application des méthodes développées pour les fonctions spéciales proprement dites au traitement des fonctions symétriques de la combinatoire algébrique. Une collaboration en commun avec M. Mishna a donné lieu à des algorithmes pour le calcul de produits scalaires entre séries de la théorie des fonctions symétriques, ce qui a permis l'énumération de familles de graphes données par des contraintes de régularité [14].

6.2.3. Élimination

Les travaux de F. Chyzak ont montré la possibilité d'algorithmiser l'intégration et la sommation définie de fonctions spéciales définies par des systèmes d'équations différentielles, aux différences ou aux q -différences via des calculs de bases de Gröbner dans des anneaux non-commutatifs. Le sujet est cependant loin d'être clos car de nombreux problèmes d'efficacité demeurent.

Dans ce contexte, B. Salvy collabore depuis quelques années avec l'équipe GAGE (École polytechnique) pour exploiter des développements récents sur l'algorithmique des *straight-line programs* de manière à produire des algorithmes et des implantations efficaces pour des problèmes de nature géométrique. L'objectif est maintenant d'étudier l'extension des méthodes fondées sur la résolution géométrique au cadre non-commutatif nécessaire pour les applications aux fonctions spéciales. Ceci fait l'objet du travail de thèse d'A. Bostan, co-encadré par M. Giusti et B. Salvy. Dans un premier temps, pour étendre la palette d'outils du cas commutatif, des algorithmes pour les éliminations faisant intervenir *deux* nombres algébriques ont été développés dans [18]. La complexité asymptotique de ces algorithmes est pratiquement optimale, à des logarithmes près.

Les bases de Gröbner interviennent par ailleurs de manière naturelle dans certains problèmes d'origine combinatoire, un joli exemple fait l'objet de [5].

6.3. Algorithmique des séquences

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

Un ensemble de recherches issues du projet (par J. Clément, P. Flajolet, P. Nicodème, B. Salvy, M. Régnier, B. Vallée) s'appuient sur la combinatoire analytique pour déterminer très précisément les probabilités d'apparition de motifs structurellement complexes. Notons qu'ainsi, au fil du temps, les travaux du projet ont permis de résoudre un très grand ensemble de problèmes portant sur la famille quasi-complète des combinaisons de contraintes suivantes : (i) un ou plusieurs motifs, voire des familles infinies de motifs : expression régulière, palindromes, ... ; (ii) une notion d'occurrence soit exacte, soit avec erreur, soit avec « trous » ; (iii) des modèles de source d'information couvrant tout aussi bien les modèles sans mémoire (Bernoulli), les modèles markoviens, voire certains modèles à mémoire infinie (sources dynamiques de B. Vallée). Ces analyses permettent notamment de construire de nombreuses formules « clefs en main » sur lesquelles s'appuyer pour

distinguer le signal du bruit dans les très nombreux problèmes d'analyses de séquences, tels qu'ils se présentent dans divers domaines de l'informatique (traitement de données textuelles, sécurité des systèmes, statistique des séquences biologiques).

Cette année, P. Flajolet, W. Szpankowski, et B. Vallée [23] ont résolu le problème de l'analyse complète en distribution de ce qu'ils appellent les « mots cachés » dans un texte. À la différence de nombreuses autres analyses en *pattern-matching*, il n'est pas imposé ici que les lettres du motif apparaissent de manière contiguë. Les formules de moyenne et variance obtenues sont aisément calculables, ce qui permet en retour de fournir des critères de décision quant à la pertinence statistique d'observations portant sur de longues séquences de symboles. La motivation initiale provient de la détection d'intrusions en sécurité des systèmes informatiques (repérage d'une suite d'événements « dangereux » qui sont « noyés » dans de grands volumes de traces), où ce problème apparaît comme basique. Cette problématique se relie à la bio-informatique dans la mesure où les motifs pertinents à l'analyse des séquences génétiques n'ont pas nécessairement leurs éléments qui apparaissent de manière ininterrompue. Ce travail est soumis au Journal of ACM.

P. Nicodème, B. Salvy et P. Flajolet ont considéré dans [9] les expressions régulières. Leur algorithme permet de calculer de manière exacte ou asymptotique l'espérance et variance du nombre d'occurrences de tout motif dans des textes de grande taille dont les lettres obéissent à une distribution de Markov. À part quelques cas de dégénérescence, la distribution limite est gaussienne. L'implantation correspondante a été réalisée par P. Nicodème en Maple, et appliquée aux statistiques d'occurrences de motifs dans les protéines.

Une autre approche du comptage des mots reliant les calculs de moyenne ou de variance à l'étude de langages particuliers a conduit à des formules explicites, obtenues par l'inversion de systèmes d'équations algébriques satisfaits par les séries génératrices. Il apparaît que le calcul effectif des formules explicites générales se simplifie lorsque les ensembles de mots ont une structure. D'un point de vue formel, on s'intéresse en particulier aux motifs approchés [24] ou structurés qui apparaissent dans de très nombreux problèmes de biologie moléculaire. Ce cas est celui des signaux de régulation des gènes, dont on connaît à priori la structure, sans connaître la séquence.

Une étude sur l'évitement des palindromes fait ainsi l'objet d'une collaboration avec M. S. Gelfand et V. Makeev (NIIGenetika). On sait qu'un site de restriction est - souvent - associé à un palindrome de longueur 4-6. L'étude vise en premier lieu à répertorier par une analyse exhaustive les palindromes présents dans les génomes bactériens, puis à valider l'hypothèse de l'évitement des palindromes associés aux systèmes de restriction-modification (SRM). En second lieu, il s'agit d'en déduire des informations sur la phylogénie de génomes bactériens proches, c'est-à-dire comparer l'utilisation des palindromes dans des génomes proches. L'analyse *in silico* basée sur la librairie *QuickScore* a permis de reproduire - et d'affiner - des résultats récents de Panina et Gelfand. Ce travail a été poursuivi [25] en autorisant des erreurs dans les motifs. M. Vandenberghe y a mis en évidence le bruit dû à une introduction incontrôlée d'erreurs, puis limité la dégénérescence des motifs aux cas autorisés par les codes d'ambiguïté IUPAC. Les résultats obtenus, mis en parallèle avec une étude phylogénétique par des programmes d'alignements multiples, de construction de matrices de distances entre séquences nucléiques d'une même classe enzymatique et de reconstruction d'arbres taxonomiques, devraient permettre de déterminer des sites de reconnaissance d'enzymes non-déterminés, et la nature et le taux de transferts horizontaux potentiels qui ont survécu dans l'évolution des génomes bactériens.

Une partie de ces fonctions a été intégrée au logiciel ScanSeq développé par V. Makeev (NIIGenetika, Moscou) et D. Papatzenko (New York University). L'application à la recherche de promoteurs impliqués dans le développement embryonnaire de la drosophile fait l'objet de [10]. L. Marsan (Université de Marne-la-Vallée) achève une thèse, encadrée par M.-F. Sagot (projet HELIX) sur la reconnaissance de signaux multiples dans les promoteurs. L'évaluation de la pertinence statistique des motifs extraits repose sur un *shuffle* de séquences et représente une des phases potentiellement coûteuses de l'algorithme. L'utilisation de formules, implantées avec M. Régner, représente un gain de temps substantiel.

M. Régner et A. Denise (Université d'Orsay) ont utilisé la théorie des grandes déviations pour obtenir les statistiques extrêmes des mots. Du fait de la structure combinatoire des mots, ils ont obtenu une expression exacte de la fonction de taux, et, tout récemment, un développement asymptotique des probabilités. Une première application est le calcul, très précis, de probabilités sur des mots, la *p*-valeur des biologistes,

lorsque les calculs exacts sont très coûteux ou numériquement instables. L'utilisation sur des génomes de plantes (thèse de Magali Lescot, Université de Marseille) a montré l'efficacité de tels calculs. Une seconde application est l'évaluation des modifications de distribution des mots induites par la sur-représentation ou sous-représentation de certains mots dans une famille de séquences. Nous avons pu obtenir des formules closes pour les espérances conditionnelles. M. Régnier et M. Vandenbogaert ont ainsi pu mettre en évidence des signaux faibles de polyadénylation masqués par un signal fort connu.

La complexité du calcul est essentiellement le coût de la résolution d'une équation polynomiale. Les résultats sont plus précis que les approximations existantes - la loi de Poisson composée utilisée dans le logiciel R'MES de l'INRA et le calcul est plus rapide que le logiciel GDon, de complexité exponentielle. Ces résultats devraient s'étendre au cas des ensembles de mots. Le cas des couples de mots est particulièrement intéressant, car il correspond à une recherche sur les deux brins de l'ADN.

L'implantation des résultats dans des procédures C optimisées a débuté dans une collaboration avec Alain Denise (Université d'Orsay). Édouard Dolley réalise un site Web, permettant l'utilisation en ligne des procédures de comptage et d'évaluation statistique écrites dans le projet. Il est prévu dans une deuxième étape, de fournir une bibliothèque de procédures C téléchargeable. Par ailleurs, certains résultats sur le calcul des p -valeurs ont été implantés dans le logiciel RSA-tools de recherche de signaux de régulation développé, à l'Université Libre de Bruxelles, par J. van Helden. Cette collaboration sera poursuivie dans le cadre du projet GenomAl.

Une collaboration avec F. Tahi et Stefan Engelen (Université d'Évry), ainsi que M. Gouy (Lyon III) a conduit à l'implantation d'un algorithme de recherche de structures secondaires, DCFold, intégrant des résultats statistiques sur les mots dans des algorithmes classiques sur les arbres de suffixes [12].

Dans le traitement de grands ensembles de données « symboliques » (au sens de grandes séquences sur un ensemble fini de symboles), deux approches sont possibles. Soit on travaille à texte variable et motif fixe : on se trouve alors principalement confronté aux problèmes évoqués ci-dessus ; soit à l'inverse, le texte est fixe - c'est la perspective des « dictionnaires », « corpus », ou grandes bases de données textuelles, sur lesquels tout prétraitement utile est légitime. Les travaux de J. Clément, P. Flajolet et B. Vallée, constituant un ensemble de 108 pages publiées en 2001, adoptent le second point de vue. Il est notamment proposé une théorie unificatrice de l'analyse des arbres digitaux (*tries*). Ces analyses sont conduites sous des modèles qui unifient la plupart des modèles de sources classiques ; sans doute, une bonne vingtaine de travaux déjà publiés sont ainsi largement étendus grâce à un cadre conceptuel simple, celui des sources dynamiques. Ce cadre détourne d'ailleurs de leur objet initial les opérateurs de transfert issus de la physique statistique et de la théorie des systèmes dynamiques en les mettant désormais au service de l'analyse d'algorithmes.

Nous collaborons sur ce thème des séquences avec d'autres projets de l'INRIA. L'approche algorithmique et combinatoire de ADAGE (G. Kucherov et G. Schaeffer), par exemple sur les mots cachés, est complémentaire de notre approche combinatoire et probabiliste. Nos résultats trouvent des applications dans les algorithmes développés par M.-F. Sagot (HELIX). L'approche probabiliste sera poursuivie avec l'action PREVAL.

7. Contrats industriels

7.1. Calcul formel

Le projet ALGO et la compagnie *Waterloo Maple Inc.* ont développé une collaboration très étroite fondée sur des intérêts réciproques. D'une part, il est intéressant pour la compagnie d'intégrer des fonctionnalités à la pointe de la recherche en calcul formel (voir la section 3.2). D'autre part, cette intégration fournit aux programmes réalisés par les membres du projet un grand nombre d'utilisateurs d'origines très diverses. Cette relation étroite nous permet également de participer aux choix effectués par les développeurs du système.

De nombreux échanges ont ainsi lieu entre le projet et la compagnie. Après une participation de plus de trois ans au projet ALGO, J. Carette a été pendant de nombreuses années *Product Development Director* à la compagnie WMI, avant de repartir cette année dans le monde académique. E. Murray, après avoir passé plus de deux ans au projet ALGO à programmer le *package* COMBSTRUCT de Maple, travaille à la compagnie WMI.

L'arrivée de Maple dans l'enseignement en classes préparatoires a nécessité un important travail de formation des enseignants. Les membres du projet ont participé activement à cet effort. C. Gomez (projet METALAU), B. Salvy et P. Zimmermann (projet SPACES) ont écrit un livre il y a cinq ans sur l'utilisation de Maple. Depuis quatre ans, P. Dumas consacre un effort important à la rédaction des solutions des exercices de ce livre et à leur mise à la disposition de tous sur le Web.

Grâce à ces nombreuses activités autour de Maple, la compagnie WMI considère l'INRIA comme un partenaire privilégié et lui accorde une licence site gratuite couvrant l'ensemble des centres. Une quinzaine de projets utilisent ce système à des degrés divers. En outre, un contrat de coopération entre WMI et le projet ALGO a été signé en 2001. Il porte en particulier sur le transfert de bibliothèques et de savoir-faire du projet dans le logiciel Maple. Dans ce cadre, les résultats d'A. Sedoglavic sont l'étude et l'implantation de calculs asymptotiques dans des échelles asymptotiques très générales. Sa bibliothèque est appelée, à terme, à remplacer toutes les routines s'occupant d'asymptotique dans le système Maple.

8. Actions régionales, nationales et internationales

8.1. Actions nationales

M. Régnier anime le projet « Algorithmique et statistique des séquences » de l'IMPG (Informatique, Mathématique et Physique pour le Génome).

Aléa est un groupe de travail dédié à l'analyse d'algorithmes et à l'analyse des propriétés des structures aléatoires discrètes. Il s'agit d'un pôle de rencontre entre informaticiens et probabilistes travaillant dans le domaine des modèles discrets. L'activité est soutenue actuellement par le groupement A.L.P. (qui encapsule l'ancien GDR/PRC A.M.I.) et est animée par P. Flajolet. L'atelier annuel s'est tenu à Marseille en mars 2002, il a rassemblé une soixantaine de participants. Par ailleurs, P. Flajolet est membre de la Commission de Reflexion sur l'Enseignement des Mathématiques (CREM, sous la direction du Pr. Kahane) et participe à ses travaux, notamment quant à l'insertion de l'algorithmique dans l'enseignement secondaire ; il est membre d'honneur de l'Association SPECIF ainsi que membre du Comité de Direction du G.D.R. A.L.P du CNRS et de diverses actions associées.

8.2. Actions financées par la Commission Européenne

Le projet ALGO est, pour une période de trois ans (2000-2003), l'une des composantes du projet ESPRIT *Long Term Research* ALCOM-FT. Ce projet rassemble dix groupes *leaders* dans le domaine de la recherche algorithmique en Europe. L'objectif affiché est la découverte de nouveaux concepts algorithmiques et l'identification des algorithmes clefs transverses à de nombreuses applications. Quatre directions de travail ont été identifiées : (i) ensembles de données massifs ; (ii) systèmes de communication complexes ; (iii) optimisation en production et planification ; (iv) recherches méthodologiques et expérimentales en algorithmique. Les travaux du projet se situent principalement dans les axes (ii) et (iv).

8.3. Relations bilatérales internationales

Une collaboration entre l'Engelhardt Institute of Computational Biology (Moscou) et le projet ALGO est soutenue depuis juin 2000 par l'Institut Liapounov.

Le projet ALGO est partenaire de l'université de Vienne et du RISC (Université de Linz) dans le cadre d'un projet Amadeus (action bilatérale France-Autriche) financé par l'Égide. Marianne Durand, lors de sa mission à Vienne, a pu collaborer avec Michael Drmota et Bernhard Gittenberger sur un problème d'urnes original, dans lequel l'urne a tendance à se vider. Ce travail a permis d'obtenir des bornes pour l'analyse en moyenne du nombre de balles dans l'urne. En visite à Linz, Ludovic Meunier a étudié avec Carsten Schneider la résolution de systèmes linéaires de récurrence avec second membre issus d'équations D-finies. Cette application des méthodes développées au RISC permet le calcul de formes closes pour les coefficients des développements en séries de certaines fonctions spéciales.

8.4. Accueils de chercheurs étrangers

Une grande partie de nos invités ont donné des exposés au séminaire du projet. Cette année, nous avons accueilli : Vsevolod Makeev, NIIGenetika, Moscou, Russie, du 6 au 14 novembre 2001 ; Helmut Prodinger, University of the Witwatersrand, Johannesburg, South Africa, du 28 novembre au 7 décembre 2001 ; Marko Petkovsek, University of Ljubljana, Slovenia, du 3 au 12 décembre 2001 ; Rafaele Giancarlo, Dipartimento di Matematica ed Applicazioni, University of Palermo, Italy, du 10 janvier au 14 février ; Volker Strehl, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany ; Alfredo Viola, Pedeciba Informatica, Montevideo, Uruguay est venu du 10 au 22 juin ; Valentina Boeva, NIIGenetika, Moscou, Russie, du 24 juin au 5 juillet ; Daniel Panario, Carleton University, Ottawa, Canada, du 1er au 5 juillet ; Marni Mishna, LACIM, Université du Québec à Montréal, Canada, du 10 septembre au 15 octobre ; Thomas Prellberg, Department of Theoretical Physics, Technische Universität Clausthal, Germany ; Hsien Kuei Hwang, Institute of Statistical Science, Taipei, Taiwan ; Oleg Vishnevsky, Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, Novossibirsk, Russie du 1er au 10 novembre ; Jacques Van Helden, Université Libre de Bruxelles, Belgique ; Peter Paule, professeur au RISC, Linz, Autriche, nous à visité une semaine, dans le cadre du projet DLMF du NIST (rédaction d'un chapitre commun).

Dans le cadre du contrat Alcom-FT, nous avons reçu Guy Louchard, Université Libre de Bruxelles, Belgique, du 29 novembre au 4 décembre et Joaquim Gabarró, Departament de Llenguatges i Sistemes Informàtics, Barcelona, Spain.

9. Diffusion des résultats

9.1. Animation de la communauté scientifique

Le projet ALGO a un séminaire régulier auquel participent plusieurs équipes partenaires de la région parisienne. Les actes en sont édités et publiés chaque année, [2].

Philippe Flajolet a été membre du comité de programme du *Eighth Seminar on Analysis of Algorithms* organisé cette année par M. Drmota à Strobl (Autriche). Cette conférence a rassemblé plus de soixante-dix spécialistes internationaux sur le sujet de l'analyse d'algorithmes. P. Flajolet est directeur du groupe de travail ALÉA, lequel se situe au sein du GDR CNRS ALP : ce groupe a tenu au CIRM (Luminy) sa réunion annuelle, laquelle a rassemblé une soixante de probabilistes, combinatoriciens, dynamiciens, et algorithmiciens du 18 au 22 mars. Il est président du comité de programme de la deuxième conférence internationale *Colloquium on Mathematics and Computer Science : Algorithms, Trees, Combinatorics and Probabilities*, (Versailles, septembre) et co-éditeur des actes publiés par Birkhäuser (560 pages), [1]. Il est éditeur de la revue *Random Structures and Algorithms*, (Wiley) et *Honorary Editor* de *Theoretical Computer Science*, (Elsevier). En 2002, il a été membre de jury ou rapporteur de diverses thèses ou habilitations. P. Flajolet est par ailleurs membre du Conseil Scientifique du GDR ALP restructuré récemment sous la coordination de C. Frougny et destiné à gérer en France une bonne partie de l'interface entre mathématiques (pures ou appliquées) et informatique fondamentale. P. Flajolet est enfin depuis 1994 Correspondant de l'Académie des Sciences (Section Sciences Mécaniques) et Membre à part entière de l'Academia Europaea depuis 1996.

Mireille Régnier a été invitée au colloque *Entangling Mathematics, Life Sciences and Information Technology : Biomimetics, a Science of multi-scale complexity* organisé par l'INTAS à Novossibirsk, Russie, visant à identifier les futures directions de recherche et actions possibles d'INTAS.

Bruno Salvy est membre du comité de pilotage de la conférence internationale de calcul formel ISSAC, en tant que représentant du groupe français de calcul formel Médiçis. Il fait partie des commissions de spécialistes de l'Université des Sciences et Technologies de Lille (en informatique) et de l'Université de La Rochelle (en mathématiques). Il est également membre du comité éditorial du *Journal of Symbolic Computation*. Cette année, il a été membre du comité de programme de la conférence ISSAC'02 et il est membre du comité de programme du Premier congrès Canada-France des sciences mathématiques qui aura lieu à Toulouse en 2004.

Il a participé aux jurys de thèse d'Olivier Ruatta (Université de Nice) et de Cyril Chabaud (Paris VI). Il a également organisé le séminaire d'évaluation du programme 2b de l'INRIA. Il maintient et anime la liste de diffusion `club-maple@inria.fr`, liste francophone consacrée au système de calcul formel Maple.

9.2. Enseignement universitaire

Frédéric Chyzak, chargé d'enseignement à temps partiel à l'École polytechnique, enseigne en tronc commun d'informatique et dans les enseignements de majeures en informatique.

Marianne Durand est enseignante à l'université de Versailles-Saint-Quentin, ou elle assure le cours d'informatique premier semestre en première année de DEUG.

Philippe Flajolet enseigne en 2002-2003 un cours sur « Méthodes asymptotiques » au DEA d'Algorithmique de la région parisienne.

Mireille Régnier enseigne un cours d'« Algorithmique et combinatoire » au DEA de bioinformatique d'Évry et un cours de Bioinformatique dans l'option « Mathématiques Appliquées » de l'École Centrale de Paris.

Mathias Vandenbogaert est vacataire à l'École Centrale de Paris. Il y a donné des travaux pratiques en bioinformatique et donné un cours sur les statistiques utilisées en bioinformatique.

9.3. Participation à des colloques, séminaires, invitations

Frédéric Chyzak a présenté ses résultats sur la manipulation de suites et fonctions spéciales par le calcul formel par des exposés et des démonstrations logicielles de son programme MGFUN au colloque *Special Functions in the Digital Age* organisé par le NIST ainsi qu'à la conférence *Foundations of Computational Mathematics*. Il a aussi donné des exposés au projet CAFE et dans les départements de mathématiques des universités d'Angers et de Rennes.

Marianne Durand a assisté et donné un exposé au séminaire AofA, à Strobl (Autriche). Elle a également rendu visite à Michael Drmota de l'université technique de Vienne, Autriche.

Philippe Flajolet a été conférencier plénier à Stanford University, à l'occasion de la « Knuthfest », ainsi qu'au Colloque Louchard de Bruxelles (janvier). Il a été invité au Colloquium de mathématiques de l'université d'Angers (février) et à celui de l'Université de Dijon (mars). Il a été également invité à l'atelier européen « RAND », dédié aux structures aléatoires et algorithmes (Paris, avril). Il a présenté les modèles de Boltzmann au *Eighth International Workshop on Analysis of Algorithms* à Strobl en Autriche, (juin) ainsi qu'au Workshop ALCOM-FT, qui s'est tenu en juillet, à Warwick, U.K. Il a été conférencier invité à l'ICM, *International Congress of Mathematicians*, à Beijing en août, où il a présenté un exposé de synthèse sur le thème *Singular Combinatorics*. Il a été invité au séminaire de théorie de l'information et ouvertures à Purdue University en octobre et a passé un mois en tant que « Visiting Scientist » sur invitation du *Department of Computer Science* de Princeton University, USA.

Ludovic Meunier a exposé un ensemble d'algorithmes pour les fonctions spéciales, lors d'une visite de 3 semaines au RISC à Linz, Autriche. Ces algorithmes sont utilisés pour produire automatiquement l'« *Encyclopedia of Special Functions* ».

Mireille Régnier a participé à la conférence RECOMB, à Washington-DC et aux journées MAS à Grenoble. Elle a présenté une communication aux journées ALÉA, au colloque BGRS'02 à Novossibirsk. Elle participe régulièrement au séminaire « Algorithmes et Biologie » de Lyon, et y a présenté une communication.

Mathias Vandenbogaert a donné un exposé dans le cadre du séminaire « Langage du Génome » de l'Université Paris-Sud. Il a participé à BBC (*3rd Belgian Bioinformatics Conference*), pour une contribution orale intitulée *Statistical measurements applied to bacterial RM-systems through genome-scale analysis, and related taxonomic issues*, [26]. Il a présenté un poster, des résultats concernant ses études effectuées sur les RMS, à JOBIM, [27]. Il a présenté un exposé intitulé *Analysis of bacterial RM-systems through genome-scale analysis and related taxonomic issues*, à BGRS, *Third International Conference on Bioinformatics of Genome Regulation and Structure* (Novosibirsk), [25]. Il a donné un exposé « Analyse de systèmes de restriction-modification bactériens par le caractère exceptionnel de leur site de reconnaissance, et inférences taxonomiques », au LaBRI à Bordeaux.

10. Bibliographie

Livres et monographies

- [1] éditeurs B. CHAUVIN, P. FLAJOLET, D. GARDY, A. MOKKADEM., *Mathematics and Computer Science II : Algorithms, Trees, Combinatorics and Probabilities*. série Trends in Mathematics, Birkhäuser Verlag, Basel, 2002, 560 pages. Proceedings of a Colloquium held at Versailles, September 2002.
- [2] *Algorithms Seminar, 2000-2001*. éditeurs F. CHYZAK., Institut National de Recherche en Informatique et en Automatique, série Research Report, volume 4406, mars, 2002, <http://www.inria.fr/rrrt/rr-4406.html>, 191 pages.

Articles et chapitres de livre

- [3] C. BANDERIER, M. BOUSQUET-MÉLOU, A. DENISE, P. FLAJOLET, D. GARDY, D. GOUYOU-BEAUCHAMPS. *Generating Functions of Generating Trees*. in « Discrete Mathematics », numéro 1-3, volume 246, mars, 2002, pages 29-55.
- [4] C. BANDERIER, P. FLAJOLET. *Basic Analytic Combinatorics of Directed Lattice Paths*. in « Theoretical Computer Science », numéro 1-2, volume 281, 2002, pages 37-80.
- [5] R. CORI, D. ROSSIN, B. SALVY. *Polynomial Ideals for Sandpiles and their Gröbner Bases*. in « Theoretical Computer Science », numéro 1, volume 276, 2002, pages 1-15.
- [6] M. DURAND. *Asymptotic Analysis of an Optimized Quicksort Algorithm*. in « Information Processing Letters », volume 85, 2003, pages 73-77.
- [7] P. FLAJOLET, K. HATZIS, S. NIKOLETSEAS, P. SPIRAKIS. *On the Robustness of Interconnections in Random Graphs : A Symbolic Approach*. in « Theoretical Computer Science », numéro 2, volume 287, 2002, pages 513-534.
- [8] P. FLAJOLET, W. SZPANKOWSKI. *Analytic Variations on Redundancy Rates of Renewal Processes*. in « IEEE Transactions on Information Theory », numéro 11, volume 48, 2002, pages 2911-2921.
- [9] P. NICODÈME, B. SALVY, P. FLAJOLET. *Motif Statistics*. in « Theoretical Computer Science », numéro 2, volume 287, 2002, pages 593-618, Extended version of an article published in the proceedings of 7th Annual European Symposium on Algorithms ESA'99, Prague, July 1999.
- [10] D. PAPTZENKO, V. MAKEEV, A. LIFANOV, M. RÉGNIER, A. NAZINA, C. DESPLAN. *Extraction of Functional Binding Sites from Unique Regulatory Regions : The Drosophila Early Developmental Enhancers*. in « Genome Research », volume 12, 2002, pages 470-481, Preliminary version in Drosophila Workshop, Washington 2001.
- [11] A. SEDOGLAVIC. *A probabilistic algorithm to test local algebraic observability in polynomial time*. in « Journal of Symbolic Computation », numéro 5, volume 33, mai, 2002, pages 735-755.
- [12] F. TAHI, M. GOUY, M. RÉGNIER. *Automatic RNA secondary structure prediction with a comparative*

approach. in « Computers and Chemistry », To appear, A preliminary version appeared as a poster at RECOMB'01 and MFRS'01 ; Montréal.

Communications à des congrès, colloques, etc.

- [13] H. BRÖNNIMANN, F. CAZALS, M. DURAND. *Randomized Jumplists : A Jump-and-Walk Dictionary DataStructure.* série Lecture Notes in Computer Science, Springer Verlag, 2002, Proceedings of the STACS'03 Conference, Berlin, February 2003.
- [14] F. CHYZAK, M. MISHNA, B. SALVY. *Effective D-finite Symmetric Functions.* in « 14th Conference of Formal Power Series and Algebraic Combinatorics », University of Melbourne, pages 19.1-19.12, Melbourne, Australia, juillet, 2002, Extended abstract. Proceedings FPSAC'02.
- [15] P. DUCHON, P. FLAJOLET, G. LOUCHARD, G. SCHAEFFER. *Random Sampling from Boltzmann Principles.* in « Automata, Languages, and Programming », série Lecture Notes in Computer Science, numéro 2380, Springer Verlag, éditeurs P. W. ET AL., pages 501-513, 2002, Proceedings of the 29th ICALP Conference, Malaga, July 2002..
- [16] P. FLAJOLET. *Singular Combinatorics.* in « Proceedings of the International Congress of Mathematicians », volume III, World Scientific, éditeurs L. T. (. DAQIAN)., pages 561-571, 2002, Invited lecture, ICM02, Beijing, China, 20-28 August 2002..
- [17] G. MATERA, A. SEDOGLAVIC. *The differential Hilbert function of a differential rational mapping can be computed in polynomial time.* in « Proceedings of the 2002 International Symposium on Symbolic and Algebraic Computation », Association for Computing Machinery, ACM press, éditeurs T. MORA., pages 184-191, Lille, France, juillet 7-10, 2002.

Rapports de recherche et publications internes

- [18] A. BOSTAN, P. FLAJOLET, B. SALVY, É. SCHOST. *Fast Computation with Two Algebraic Numbers.* Research Report, numéro 4579, Institut National de Recherche en Informatique et en Automatique, octobre, 2002, <http://www.inria.fr/rrrt/rr-4579.html>, 20 pages.
- [19] P. FLAJOLET, B. SALVY, G. SCHAEFFER. *Airy Phenomena and Analytic Combinatorics of Connected Graphs.* Research Report, numéro 4581, Institut National de Recherche en Informatique et en Automatique, octobre, 2002, <http://www.inria.fr/rrrt/rr-4581.html>, 25 pages.
- [20] P. FLAJOLET, R. SEDGEWICK. *Analytic Combinatorics-Symbolic Combinatorics.* Research Report, numéro 4103, INRIA, 2002, <http://www.inria.fr/rrrt/rr-4103.html>, 186+vii pages.
- [21] B. SALVY, J. SHACKELL. *Asymptotic Expansions with Oscillating Coefficients.* Research Report, numéro UKC/IMS/02/33, University of Kent, Canterbury, octobre, 2002.

Divers

- [22] J. FAYOLLE. *Paramètres des arbres suffixes dans le cas de sources simples.* 2002, Mémoire de DEA, Université Paris VI..

-
- [23] P. FLAJOLET, W. SZPANKOWSKI, B. VALLÉE. *Hidden Word Statistics*. Preprint, 2002, 34 pages ; submitted to *J. of the ACM*.
- [24] M. RÉGNIER. *Mathematical tools for regulatory signals extraction*. 2002, presented at BGRS'02 ; submitted.
- [25] M. VANDENBOGAERT, V. MAKEEV. *Analysis of bacterial RM-systems through genome-scale analysis and related taxonomic issues*. 2002, presented to BGRS'02, Novossibirsk.
- [26] M. VANDENBOGAERT. *Statistical measurements applied to bacterial RM-systems through genome-scale analysis, and related taxonomic issues*. 2002, presented at BBC 2002, 3rd Belgian Bioinformatics Conference.
- [27] M. VANDENBOGAERT. *Statistical measurements applied to bacterial RM-systems through genome-scale analysis*. 2002, presented to JOBIM 2002.