INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet ALGO

Algorithmes

Rocquencourt

THÈME 2B

Table des matières

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]
Philippe Robert [DR]
Vincent Dumas [MC, Université de Bordeaux, en délégation]

Collaborateurs extérieurs
Cyril Banderier [Bourse Post-doc INRIA, Sarrebruck, à partir du 1er octobre]
Julien Clément [CR CNRS, Université de Caen]
Philippe Dumas [professeur, Cl. Prépa. lycée Jean-Baptiste Say]
Alain Dupuis [France Télécom R&D, Lannion]
Fabrice Guillemin [France Télécom R&D, Lannion]
Pierre Nicodème [CR CNRS, Université d'Évry]
Maryse Pelletier [professeur, Université Paris VI]
Mathieu Raffinot [CR CNRS, Université d'Évry]
Michèle Soria [professeur, Université Paris VI]
Brigitte Vallée [professeur, Université de Caen]

Chercheurs Invités
Marni Mishna [Université du Québec à Montréal, du 12 septembre au 31 décembre]
Michael Shalmon [INRS-Télécommunications, Université du Québec à Montréal, Canada, du 1er septembre 2000 au 28 février]

Chercheurs post-doctorants
Alexandre Sedoglavic [à partir du 1er octobre]
Albertus Zwart [à partir du 1er octobre]

Doctorants
Cyril Banderier [Université de Paris VI, jusqu'au 30 septembre]
Alin Bostan [École polytechnique, co-encadrement]
Marianne Durand [ENS Paris]
Mostapha Haddani [Université de Versailles St-Quentin]
Grégoire Lecerf [École polytechnique, jusqu'au 14 septembre, co-encadrement]
Ludovic Meunier [École polytechnique]
Vincent Puyhaubert [ENS Cachan, à partir du 3 septembre]
Mathias Vandenbogaert [Université de Bordeaux]

Stagiaires
Jitesh Jain [Indian Institute of Technology, Kanpur, Inde, du 10 mai au 22 juillet]
Thomas Klausner [Technische Universität Wien, Austria, du 15 janvier au 15 avril]

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'optimisation de l'allocation de ressources dans les réseaux de communication. La réalisation de cet objectif passe par la compréhension en profondeur de l'aléa 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.

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 « aléatoirisation », 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), les protocoles en arbre (P. Flajolet et P. Jacquet), et le comportement probabiliste de grands réseaux (P. Robert). 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, Newcastle, Princeton, Purdue, Stanford, Vienne, Washington, Waterloo, etc.

3 Fondements scientifiques

3.1 Analyse d'algorithmes

Participants : Cyril Banderier, Julien Clément, 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.

Résumé : *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 (voir la section 6.1) ; 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.

Résumé : *Le projet* Algo *développe des algorithmes de calcul formel qui permettent : le traitement de modèles combinatoires récursifs ; 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 Saga, 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.

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 spé-

cification ; *(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 fonctionalités 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 specification, 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).

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). 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.

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 est en cours d'intégration 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 : Frédéric Chyzak, Philippe Flajolet, Jitesh Jain, Pierre Nicodème, Mireille Régnier, Bruno Salvy, Mathias Vandenbogaert.

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

Résumé : *L'objet des recherches sur les séquences est la conception de nouveaux algorithmes, l'obtention de la complexité moyenne de ces algorithmes, l'implémentation 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-representés ou

sous-representés 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-representation 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 precision 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.

3.4 Algorithmique et modélisation des réseaux

Participants : Jean-François Dantzer, Vincent Dumas, Mostafa Haddani, Philippe Robert.

Mots clés : grands réseaux, méthode de renormalisation, vitesse de convergence.

Résumé : *L'étude quantitative des modèles probabilistes de réseaux, complexe dès lors que le nombre de nœuds est grand, est envisagée du point de vue des techniques de renormalisation de la physique statistique. L'état du réseau est renormalisé avec un paramètre (la taille, la capacité du réseau, l'intensité du trafic) qui tend vers une valeur critique. Asymptotiquement, l'étude du comportement du réseau se ramène à la résolution d'équations différentielles quasi-déterministes.*

Le cadre général de cette recherche concerne les propriétés de renormalisation des réseaux de communication. Le processus de Markov décrivant l'état d'un réseau est en général complexe, même si la distribution à l'équilibre de celui-ci est connue. La combinatoire des expressions ne permet pas une évaluation qualitative de ces réseaux, pour les problèmes de dimensionnement notamment, lorsque leur nombre de nœuds est significatif. Une méthode intéressante, issue de la physique des particules, consiste à renormaliser le processus à la fois en temps et en espace par un petit paramètre ε et faire tendre celui-ci vers 0. Un processus limite ainsi obtenu conserve les caractéristiques essentielles du réseau, schématiquement la partie bruit autour des trajectoires est éliminée. Le processus limite obtenu de cette façon est quasi-déterministe, les

points de discontinuité de la dynamique conservant une part d'aléatoire. Toute la difficulté de l'étude consiste à identifier et caractériser les limites possibles. Il peut y avoir plusieurs limites et les équations « limites » peuvent présenter des solutions pathologiques qu'il convient d'éliminer (les physiciens le font avec des arguments d'entropie). La renormalisation présente l'intérêt de mettre en évidence les modes de fonctionnement fondamentaux. Ce programme d'étude a déjà largement été entamé dans la thèse de V. Dumas et développé par J.-F. Dantzer dans le cadre des réseaux multi-classe et par M. Haddani dans une thèse en cours sur les questions d'allocation de bande passante.

Les problèmes de vitesse de convergence constituent un aspect important de notre étude. Ils se situent en amont des questions de renormalisation. Schématiquement, l'étude d'un système donné peut se décrire de la façon suivante : étude de la renormalisation, propriétés gaussiennes autour des trajectoires renormalisées, existence d'un principe de grandes déviations, et enfin étude de la convergence à l'équilibre des modèles considérés. Jusqu'à présent, les résultats obtenus dans ce domaine sont principalement des estimations du taux de convergence exponentielle par rapport au temps, pour obtenir des bornes du type $Ke^{-\lambda_2 t}$. La constante K, qui est généralement négligée, joue cependant assez souvent un rôle majeur dans la convergence à l'équilibre. Par exemple, l'étude de la seconde valeur propre λ_2 suggère souvent un temps de relaxation (le temps pour que la distance à l'équilibre soit inférieur à $1/e$) de l'ordre de $1/\lambda_2$. Il est cependant très fréquent que l'ordre de grandeur du véritable temps de relaxation soit complètement différent. L'accumulation de valeurs propres au voisinage de λ_2 semble être à l'origine de ces phénomènes encore mal compris. Notre objectif principal dans ce cadre est de donner des *bornes* aussi simples et précises que possible sur la distance entre l'état d'un système à un instant donné et ce même système à l'équilibre. L'intérêt de tels résultats est de pouvoir donner une description du comportement transitoire d'un système qui n'est pas à l'équilibre. Si l'état d'équilibre a, en général, une expression difficile à utiliser, le cas transitoire est en revanche pratiquement toujours inconnu. Ce type d'étude a été mené avec succès, avec des techniques d'analyse de Fourier. Les marches aléatoires sur des graphes ayant une structure assez régulière s'étudient avec ces méthodes, voir par exemple le livre de Diaconis « *Group representations in probability and statistics* » sur ce sujet. Les techniques envisagées pour cette étude sont essentiellement probabilistes. Notre objectif est d'étudier dans un premier temps des réseaux simples et de dégager des méthodes d'investigation. Dans ce cadre aussi, nous avons recours à une variable de renormalisation qui permet de dégager le mode majeur de convergence. Outre les bornes explicites sur la distance à l'équilibre, ce type d'étude permet de définir, dans certaines situations, le temps d'atteinte de l'équilibre ; avant ce temps, il est possible de trouver un état initial pour lequel la distance est maximale, et après la distance est nulle.

4 Domaines d'applications

4.1 Panorama

Résumé : *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, le traitement statistique des séquences biologiques, et les protocoles de communication.

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.

Les travaux sur les réseaux sont naturellement liés aux problèmes de dimensionnement, de routages des messages, ou encore d'optimisation des paramètres de qualité de service (délai, taux de rejet, etc.). Les grandes tailles des réseaux, les hauts débits utilisés et l'hétérogénéité des trafics (vidéo, voix, données) qui y circulent rendent de plus en plus nécessaire une utilisation accrue de modèles mathématiques.

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. 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 plus d'une cinquantaine de nouvelles fonctions à l'utilisateur. 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 maintenant intégré en une unique bibliothèque du nom d'ALGOLIB, et diffusée par les pages web du projet ALGO. Leur installation par les utilisateurs en est ainsi grandement simplifiée, tout en permettant une meilleure interaction entre les *packages*.

6 Résultats nouveaux

6.1 Analyse d'algorithmes

Participants : Cyril Banderier, Julien Clément, Marianne Durand, Philippe Flajolet,

Mireille Régnier, Bruno Salvy, Michèle Soria, Brigitte Vallée.

Les recherches en 2001 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 la section 6.4). 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.

Combinatoire analytique Un premier lot de travaux conçus ou publiés en l'an 2001 a trait 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 concrétisé par un rapport de 98 pages [34] a consisté à dégager un noyau de propriétés essentielles à l'analyse de tous les modèles qui s'expriment par des fonctions rationnelles ou algébriques. Ceci recouvre le comportement quantitatif de tous les modèles d'états finis (utilisés tant en algorithmique textuelle, qu'en modélisation de systèmes informatiques ou en vérification) et l'ensemble des objets susceptibles d'être décrits par des grammaires dites *context-free* (couvrant par exemple de nombreuses configurations géométriques planes, les structures secondaires d'ARN, ou encore des problèmes variés d'analyse syntaxique). Le plan de la « combinatoire rationnelle » rejoint naturellement la théorie de Perron-Frobenius des opérateurs positifs, elle-même à la base des chaînes de Markov en dimension finie. Cette « combinatoire rationnelle » inclut par ailleurs la plupart des problèmes liés à la recherche de motifs dans les séquences, pour lesquels elle constitue un cadre mathématique naturel (Section 6.3). La « combinatoire algébrique » a été jusqu'ici largement traîtée de manière *ad hoc*, au gré des problèmes individuels. L'intérêt principal de [34] est de proposer pour la première fois un ensemble de procédures de décision permettant de calculer automatiquement le comportement asymptotique des coefficients de toute fonction algébrique. La finalisation de cette question se fera en liaison directe avec l'axe « calcul formel » du projet au cours de l'an 2002, puisque l'on peut déjà envisager l'implantation en MAPLE de procédures de décision complètes.

Poussée par l'objectif de procéder à une classification des principaux schémas de la combinatoire analytique, une étude de fond par C. Banderier, P. Flajolet, G. Schaeffer, et M. Soria [4] a été consacrée à dégager une classe d'universalité (terme issu de la physique statistique) en combinatoire analytique. Il s'agit d'analyser des « transitions de phase » dont on ressent de plus en plus en plus la présence par exemple dans les phénomènes de seuil observés en optimisation combinatoire (avec régions « faciles » et « difficiles » séparées brusquement). Jusqu'ici, de nombreux phénomènes échappaient à la quantification. Il a été mis en évidence au sein

du projet le rôle important de la coalescence de cols et de la confluence de singularités qui conduisent à des phénomènes décrits par fonctions d'Airy très différents des habituels comportements gaussiens. Notons d'ailleurs que ces phénomènes se retrouvent dans des domaines très divers,tels la coalescence dans les tables de hachage [6, 15], ou la connectivité dans les graphes aléatoires. Depuis l'automne 2001, un nouveau doctorant, V. Puyhaubert explore l'impact possible de telles méthodes analytiques sur les instances de problèmes « difficiles », c'est-à-dire, NP-complets (partitionnement entier, satisfaisabilité de formules booléennes, par exemple).

Un autre ensemble de travaux est relatif aux chemins aléatoires, objets de base de la combinatoire et des probabilités discrètes. Il s'agit là de l'un des thèmes principaux développés dans la thèse de C. Banderier [2] soutenue en juin 2001. L'intérêt des chemins plans est de coder naturellement certaines des structures de base de l'informatique, comme les arbres ou les traces d'évolution d'un système informatique dont la taille varie dynamiquement. Dans le même temps, les chemins aléatoires donnent lieu à un processus stochastique fondamental, le mouvement brownien, de sorte qu'ils se situent à un intéressant carrefour entre informatique théorique, combinatoire, et théorie des probabilités. L'article [5] développe notamment à partir de la thèse de C. Banderier [2] une théorie asymptotique unifiée des chemins à ensembles finis de sauts permis.

Ces études ont par ailleurs des répercussions en génération (aléatoire) de jeux de tests : ici, l'on parvient, dans le prolongement de travaux de G. Schaeffer, à engendrer rapidement des graphes planaires (« cartes ») vérifiant des contraintes de connectivité d'ordre élevé. Les méthodes de génération aléatoire induites par ces études sont ensuite susceptibles de servir à la génération automatique de jeux de tests en optimisation combinatoire ou en dessin de graphes. Certaines de ces pistes sont d'ailleurs actuellement explorées en collaboration avec le Max Planck Institut für Informatik (MPI) de Saarbrücken où C. Banderier effectue un séjour postdoctoral d'un an à partir d'octobre 2001. Elles se situent notamment dans l'axe « algorithmique expérimentale » du projet Européen ALCOM-FT dont ALGO et MPI sont partie prenante.

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.

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 l'algorithme de tri généraliste utilisé par excellence dans de nombreux systèmes informatiques, à cause notamment 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 [41] 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

optimisé par Bentley sur la base empirique de simulations et pour laquelle l'analyse apporte une validation théorique plaisante, à quelques pour cent près. M. Durand a également résolu avec S. Taylor [12] les récurrences qui apparaissent dans un modèle approché d'évolution du coût de la recherche d'un élément dans un arbre binaire de recherche lorsque celui-ci est soumis à des insertions-suppressions successives. Par ailleurs, une collaboration en cours avec Michael Fellows porte sur des récurrences (définies par cas sur des régions disjointes du plan) qui sont motivées par la problématique générale de la « complexité paramétrée ».

La gestion des tables de hachage est un autre domaine ancien de l'analyse d'algorithmes (la première analyse de D. Knuth vers 1963 y est relative, cf. [16]). Ces classes de méthodes étant fondées sur « l'aléatoirisation » (*randomization*), elles ne sont par principe validées que par l'analyse probabiliste. Les deux articles [6, 15] font de fait le point des résultats récents et proposent une caractérisation fine de la distribution asymptotique des coûts. La distribution ainsi mise en lumière se relie à des questions multiples : la connectivité des graphes aléatoires, l'aire sous une excursion brownienne, la longueur de cheminement dans les arbres aléatoires, etc.

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 par exemple le rapport d'activité du Projet HIPERCOM à Rocquencourt, dans les dernières années. Une théorie globale de l'analyse probabiliste des arbres digitaux fondée sur les « sources dynamiques » de B. Vallée est détaillée dans une étude de 63 pages parue en 2001 dans la revue *Algorithmica*, et discutée à la section 6.3.

Sur un registre autre, la factorisation de polynômes intervient tant dans la conception de nombreux codes correcteurs d'erreur (cf. le rapport d'activité du Projet CODES à l'INRIA) que dans de nombreux domaines liés au calcul formel (intégration formelle, par ex.). La chaîne classique dite de Berlekamp-Kantor-Zassenhaus reçoit pour la première fois une analyse complète donnée dans [13]. Cet article est le résultat de la coopération entre P. Flajolet, X. Gourdon (ancien doctorant du Projet, actuellement ingénieur à Dassault-Systèmes) et D. Panario (Ottawa). Il présente l'intérêt d'illustrer le fait que les méthodes de combinatoire analytique ne se limitent pas aux objets classiques de la combinatoire mais s'étendent à de nombreux domaines de l'algorithmique probabiliste.

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 [38] 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 2002 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 test. 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 Bruxelles.

6.2 Calcul formel

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

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 dépendence de ce comportement vis-à-vis des diverses variables. De nombreuses questions liées à une classification complète des comportements singulier 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, qui poursuit maintenant une thèse à la Technische Universität Wien, en Autriche.

Son travail prolonge la réalisation par M. Mishna que les grammaires attribuées fournissent un bon cadre pour exprimer les algorithmes dont la complexité est susceptible d'être analysée automatiquement [29]. Cette idée est désormais implantée en COMBSTRUCT, qui produit ainsi automatiquement des équations de séries génératrices multivariées.

Dans le cas univarié, les outils asymptotiques développés depuis de nombreuses années par B. Salvy permettent désormais d'aborder des calculs que leur complexité rend difficilement amenable à un traitement manuel. Ainsi [35] répond à un problème posé par H. Wilf concernant l'étude asymptotique de la différence entre le nombre de partitions d'ensembles à nombre de parts pairs et à nombre de parts impaires. Ceci représente un problème asymptotique délicat : du point de vue analytique, deux points cols conjugués dictent le comportement, mais il faut obtenir cinq ou six termes du développement asymptotique pour avoir un bon accord avec les valeurs exactes.

Fonctions spéciales L. Meunier [28] a développé 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 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. 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.

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. Ces travaux ont abouti à un nouvel algorithme de résolution de systèmes polynomiaux qui a fait l'objet d'un article d'une cinquantaine de pages paru cette année [17] et d'une partie de la thèse de G. Lecerf, soutenue en septembre 2001, co-encadré par M. Giusti (CNRS) et B. Salvy. L'implantation prototype de G. Lecerf, en MAGMA, permet de résoudre exactement des systèmes polynomiaux de grande taille de manière très compétitive par rapport aux meilleurs logiciels disponibles pour les calculs de bases de Gröbner.

Les systèmes polynomiaux interviennent également dans une nouvelle construction pour des ondelettes d'un certain type [7] dans un travail mené en collaboration par F. Chyzak, avec P. Paule et B. Zimmermann du RISC (Linz, Autriche), O. Scherzer et A. Schoisswohl de l'*Industrial Mathematics Institute* (autre équipe de l'université de Linz). Une étape cruciale de la construction est la détermination sous forme explicite par le calcul formel d'une famille de coefficients dits « de filtrage » qui représentent de façon compacte toute l'information d'une ondelette. Ces coefficients sont donnés comme solutions de systèmes d'équations polynomiales, dont les bases de Gröbner ont permis la résolution.

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 pallette d'outils du cas commutatif, un algorithme accéléré de calcul de polynômes minimaux dans des algèbres quotient a été mis au point [33].

6.3 Algorithmique des séquences

Participants : Philippe Flajolet, Jitesh Jain, 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'appuyent 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 [22]). 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 domaine de l'informatique (traitement de données textuelles, sécurité des systèmes, statistique des séquences biologiques).

Cette année, P. Flajolet, Y. Guivarc'h, W. Szpankowski, et B. Vallée [27] 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 de ce problème 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 a été accepté et présenté à l'ICALP-2001.

Notre 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 s'ensuit [31] que la complexité du calcul dépend de la taille minimale d'un automate de « recouvrement ». Cependant, 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 expressions régulières et aux motifs approchés [44] qui apparaissent dans de très nombreux problèmes de biologie moléculaire. Les formules obtenues pour différentes structures d'ensembles de mots [44] ont été implémentées en C dans la bibliothèque *QuickScore* dans le cadre d'un stage de J. Jain, prolongeant un travail commencé dans le cadre de l'Action REMAG coordonnée par G. Kucherov (Projet ADAGE).

Une partie de ces fontions a été integrée au logiciel ScanSeq developpé par V. Makeev (NIIGenetika, Moscou) et D. Papatzenko (New York University). L'application à la recherche de promoteurs impliqués dans le developpement embryonnaire de la drosophile fait l'objet de [20]. 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, implémentées avec M. Régnier, représente un gain de temps substantiel.

Enfin, une étude sur l'évitement des palindromes fait l'objet d'une collaboration avec M. S. Gelfand et V. Makeev (NIIGenetika). Elle 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. 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 se poursuit en autorisant des erreurs dans les motifs, c'est-à-dire en considérant des motifs dégénerés, selon les codes d'ambiguité IUPAC.

M. Régnier et A. Denise (Université d'Orsay) ont utilisé la théorie des grandes déviations pour obtenir les statistiques extrêmes des mots. Une première application est le calcul, très précis, de probabilités sur des mots, lorsque les calculs exacts sont très coûteux ou numériquement instables. Une comparaison avec des méthodes existantes a été presentée à WABI'01 [24]. Cette approche permet aussi d'obtenir des formules closes pour les espérances conditionnelles. Il s'agit ici d'évaluer comment la sur-représentation de certains mots dans une famille de séquences modifient la distribution des autres mots.

Une collaboration avec F. Tahi (Université d'Évry) et M. Gouy (Lyon III) a conduit à l'implémentation 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 [21].

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 [8, 22], 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* en anglais). 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 developpés par M.-F. Sagot (HELIX).

6.4 Algorithmique et modélisation des réseaux

Participants : Vincent Dumas, Alain Dupuis, Mostapha Haddani, Fabrice Guillemin,

Philippe Robert, Bert Zwart.

Étude microscopique de TCP Cette partie concerne l'étude de la transmission des paquets par une connexion TCP acheminant un fichier de taille infinie à travers le réseau [25, 40]. Plus précisément, nous avons étudié le comportement asymptotique de la taille de la fenêtre de congestion associée à la connexion. En utilisant un modèle très simple où le réseau perd des paquets de façon aléatoire, nous avons montré que la suite des carrés des tailles des fenêtres de congestion convenablement renormalisées formait une suite auto-régressive. Ce résultat est, à notre connaissance, le premier qui mette en évidence cette remarquable propriété.

Comme conséquence, cela permet d'obtenir facilement les expressions explicites du débit de la connexion TCP ainsi que la densité de la distribution à l'équilibre de la taille de la fenêtre de congestion. La queue de distribution de celle-ci décroît en $\exp(-ax^2)$ et non exponentiellement comme pouvaient le suggérer certains modèles utilisés précédemment dans la littérature. Qualitativement, cela implique que le nombre de grandes fenêtres de congestion est surestimé par ces modèles. Il faut noter que nos résultats intègrent le fait que le protocole a une fenêtre de congestion de taille maximale, ce qui n'est pas pris en compte par la plupart des modèles actuels. Les expressions analytiques des résultats sont sensiblement plus compliquées par rapport au cas de la taille infinie de la fenêtre de congestion maximale.

Le comportement transitoire du protocole a aussi été étudié, en particulier le temps nécessaire pour atteindre la fenêtre de congestion maximale (*i.e.* le débit maximum). Pour cette variable, un résultat limite a été démontré. Il est intéressant de noter que la valeur de sa moyenne s'exprime à l'aide d'une fonction définie sur une structure arborescente de $\mathbf{R}^2$.

Actuellement, plusieurs directions d'étude sont envisagées :

- la généralisation de ces résultats aux processus de perte observés sur Internet (le modèle considéré ici étant une sous-classe de ces processus de perte). Les modèles mis en évidence par les mesures de Paxson sont au cœur de cette étude.
- La validation des résultats en utilisant le simulateur NS. Le protocole TCP dans notre étude est réduit à sa partie *congestion avoidance*, les algorithmes *slow start*, *fast recovery* et *fast retransmit* étant négligés. Il est possible de montrer que lorsque le taux de perte tend vers 0, cette approximation est valide. Les simulations montrent que pour des taux de perte de l'ordre de 10^{-4} cette approximation est valide et que la queue de distribution décroît bien comme $\exp(-ax^2)$.
- Étude de « vraies » traces TCP. (Voir la section sur le projet Métropolis).

Étude macroscopique de TCP Une proposition d'étude de l'allocation de bande passante dans les réseaux a été acceptée dans le cadre des consultations thématiques de France Télécom R&D. Cette étude se fait en collaboration avec A. Dupuis et F. Guillemin (Lannion).

Elle se propose d'investiguer les problèmes d'allocation de bande passante dans des réseaux transportant du trafic élastique, comme par exemple le trafic TCP dans le réseau Internet. Actuellement dans un réseau de télécommunication les messages se partagent la bande passante de façon égalitaire. Par exemple, l'implémentation de TCP est telle que les ajustements se font sur les nœuds les plus chargés, à ces nœuds la bande passante est équitablement répartie entre les messages. Si les mécanismes de ce type de politique ont l'avantage de réguler correctement,

de façon distribuée le trafic, ils présentent l'inconvénient de ne pas utiliser pleinement la capacité du réseau. En effet, en prenant un cas simplifié, si un message passe par une série de N nœuds peu chargés sauf un où passent M messages, en supposant que la bande passante maximum soit λ à chaque nœud, le protocole TCP fera que le message sera transmis au taux λ/M à travers le réseau. Globalement une petite fraction de la capacité totale du réseau sera utilisée λ/M par nœud au lieu de λ dans le cas idéal.

Il est donc intéressant d'étudier la possibilité d'augmenter l'utilisation de la capacité d'un réseau en modifiant les algorithmes de partage de bande passante à chaque nœud tout en préservant certaines des caractéristiques actuelles du protocole : les mécanismes doivent être distribués et adaptatifs. C'est le but principal de l'étude menée avec France Télécom R&D.

Dans un premier temps une étude détaillée de la politique d'allocation de bande passante par TCP est conduite. Cette politique qui est approximée par une politique appelée Max-min est complexe à étudier qualitativement (sa description algorithmique n'est d'ailleurs pas complètement triviale).

Cette année nous avons étudié le cas d'un réseau linéaire ayant une capacité maximale C où des connexions permanentes sont établies entre deux points pris au hasard sur le réseau [26]. Dans ce cadre nous avons considéré la politique « Min » qui alloue la bande passante C/N à une connexion qui partage une section du réseau avec au maximum $N-1$ autres connexions. Cette politique présente l'avantage de sous-estimer la bande passante effectivement attribuée à une connexion par la politique Maxmin. Dans le cadre du réseau linéaire la politique Min s'exprime à l'aide de la file d'attente $M/M/\infty$. Nous avons montré que la distribution de la bande passante inutilisée pour cet algorithme s'exprime à l'aide des polynômes orthogonaux de Poisson-Charlier. Quand le nombre de connexions devient très grand, nous avons montré que la fraction de la bande passante inutilisée tend vers 0. Le taux de décroissance vers 0 en fonction de la charge a aussi été explicité. Pour ces réseaux, la politique Maxmin n'est donc pas pénalisante puisque la bande passante inutilisée est négligeable. D'autres topologies sont actuellement à l'étude ainsi que les aspects dynamiques (très délicats) de ces politiques. Ces sujets sont traités par M. Haddani dans sa thèse. Cette dernière doit être soutenue à la fin de l'année.

Métropolis Le projet RNRT « Métropolis » qui commence le 1er septembre 2001 réunit les deux volets de cette étude (Macroscopique vs microscopique). Ce projet regroupe le département réseau du LIP6, l'Institut Eurecom, France Télécom R&D, Groupe des Écoles de Télécommunications (GET), le LAAS, RENATER et l'INRIA. Pendant ce projet des expériences seront menées sur le trafic IP sur plusieurs sections du réseau RENATER entre les centres de Lannion, Paris, Toulouse et Nice. Des mesures très précises flot par flot seront effectuées pour ensuite pouvoir discriminer le trafic global : trafic « lourd » (*elephants*) ou léger (*mice*), détecter les engorgements locaux, donner les statistiques des processus de perte (notamment caractériser la distribution de la taille des groupes de paquets perdus en cas de congestion), impact de la phase de slowstart, etc... De plus, la validation de résultats obtenus dans [25] peut être envisagée en utilisant cette campagne intensive de mesures sur le réseau. Il faut noter que les mesures disponibles actuellement sur le réseau n'ont pas le degré de précision de celles qui seront effectuées dans Métropolis.

Contrôle d'admission dans une file d'attente multiclasse avec priorité Nous proposons un nouvel algorithme de contrôle d'admission [39] dans le cadre d'un modèle qui domine (au sens du taux de perte) le modèle classique de trafics régulés par des *leaky-buckets*. Dans notre modèle, les paquets arrivent en *batches* (groupés) selon un processus de Poisson et demandent un temps de service exponentiel à un unique serveur, les paquets en attente étant stockés dans un *buffer* de taille K (avec perte si un paquet trouve le *buffer* plein). Différentes classes de trafics coexistent, une classe étant caractérisée par l'intensité du processus de Poisson et par la taille des *batches*. La question est de déterminer combien de trafics de chaque classe peuvent être admis en maintenant le taux de perte en deçà du seuil autorisé.

L'intérêt de ce modèle est qu'une analyse exacte peut être conduite, y compris lorsqu'on y ajoute un trafic prioritaire (les questions de priorités s'annonçant comme primordiaux dans les réseaux de demain). Dans ce cas, la règle empirique suivie jusqu'à présent (*Reduced Service Rate*) consistait à réduire en conséquence le taux de service des trafics non prioritaires. Notre analyse montre que la RSR n'est pas une règle fiable et conservative.

La propriété remarquable, dans les deux cas (avec ou sans trafic prioritaire), est que les régions d'admissibilité sont délimitées par des hyperplans. Chaque classe de trafic se voit ainsi affecter une constante explicite (communément appelée *effective bandwidth*, ou bande passante effective), lesquelles constantes suffisent à définir l'algorithme, et ont l'avantage (en l'absence de trafic prioritaire) de ne dépendre que de la classe à laquelle elles sont affectées.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Algorithmique et modélisation des réseaux

M. Haddani et P. Robert participent à la consultation thématique de France Télécom R&D sur l'optimisation de la gestion du trafic TCP dans un réseau. (Voir la partie résultats nouveaux pour la description de cette action). Ce contrat est d'une durée de deux ans.

V. Dumas et P. Robert participent au projet RNRT Métropolis sur l'utilisation de la métrologie dans l'étude des réseaux IP (Voir la description dans la section résultats nouveaux). Ce contrat est d'une durée de trois ans.

7.2 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. J. Carette est retourné à la compagnie WMI, après une participation de plus de trois ans au projet ALGO. Il y est maintenant *Product Development Director* et s'occupe de tous les aspects techniques de Maple. De même, E. Murray, après avoir passé plus de deux ans au projet ALGO à programmer le *package* COMBSTRUCT de Maple, travaille maintenant 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 trois 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 cette année. Il porte en particulier sur le transfert de bibliothèques et de savoir-faire du projet dans le logiciel Maple. Dans ce cadre, A. Sedoglavic a démarré un post-doc au sein du projet en octobre, avec pour but l'étude et l'implantation de calculs asymptotiques dans des échelles asymptotiques très générales.

8 Actions régionales, nationales et internationales

8.1 Actions nationales

M. Régnier a participé à l'Action de Recherche Coopérative REMAG rassemblant des chercherus de l'IRISA, de l'Institut Pasteur et d'Orsay. L'objectif est la recherche de motifs dans le génome. Elle anime le projet « Algorithmique et statistique des séquences » de l'IMPG.

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 a Marseille, il a rassemblé une cinquantaine de participants.

Une liste de diffusion des utilisateurs francophones de Maple a été créé il y a cinq ans. Cette liste et son archive web sont hébergés par le projet ALGO.

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. Qautre directions de travail ont été identifiées : (*i*) ensembles de données masifs ; (*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*).

Le projet ALGO pilote la composante française du projet INTAS 99-1476, *Methods, algorithms and software for functional and structural annotation of complete genomes*, d'une durée de dix-huit mois. Ce projet implique quatre autres équipes (russe, allemandes (EMBL et MIPS) et autrichienne).

8.3 Relations bilatérales internationales

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

Dans le cadre du Programme Conjoint de Recherche France/Hong-Kong, intitulé PROCORE, nous avons une collaboration avec Mordecai Golin de l'université HKUST à Hong-Kong.

8.4 Accueils de chercheurs étrangers

Une grande partie de nos invités ont fait des exposés au séminaire du projet. Cette année, nous avons accueilli : Donald Lutz, San Diego State University, California, U.S.A, invité de l'Université d'Angers ; Vijay Vazirani, Georgia Institute of Technology, Atlanta, U.S.A. ; Alban Quadrat, School of Mathematics, University of Leeds, Royaume-Uni ; Wojciech Szpankowski, Purdue University, W. Lafayette, U.S.A. nous a rendu 2 visites ; Burkhard Zimmermann, RISC, Linz, Autriche, est venu du 13 au 28 janvier ; Thomas Klausner, Technische Universität Wien, Autriche, a passé 3 mois au sein du projet à partir du 15 janvier ; Arnold Schönhage, Institut für Informatik II, Universität Bonn, Allemagne ; Andreas Weiermann et son étudiant, Gyesik Lee, Department of Mathematics and Computer Science, University of Münster, Allemagne ; Yoshiaki Itoh, The Institute of Statistical Mathematics, Tokyo, Japon ; John Shackell, Institute of Mathematics and Statistics, University of Kent at Canterbury, Royaume-Uni, du 2 au 14 septembre ; Marni Mishna, LaCIM, Université du Québec à Montréal, Canada, du 12 septembre au 31 décembre ; Bert Zwart, Department of Mathematics and Computer Science, Eindhoven University of Technology, Pays-Bas ; Jonathan Borwein, Department of Mathematics and Statistics, SFU, Vancouver, B.C. Canada, nous a rendu visite en octobre, visite au cours de laquelle et il a été conférencié du Colloquium de Rocquencourt.

Dans le cadre du contrat Alcom-FT, nous avons reçu : Paul G. Spirakis, CTI Patras, Grèce, du 20 au 22 novembre ; Amir Dembo, Maths and Statistics Dept., Stanford University, U.S.A. ; Renzo Pinzani, Università degli Studi di Firenze, Dipartimento di Sistemi e Informatica, Italie ; Guy Louchard, Université Libre de Bruxelles, Belgique ; Michael Drmota, Institut für Geometrie, TU Wien, Autriche ; Donatella Merlini, Università di Firenze, Italie.

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

Le projet Algo a un séminaire régulier auquel participent plusieures équipes partenaires de la région parisienne. Les actes en sont édités et publiés chaque année [1].

Cyril Banderier s'est occupé de l'Association des Doctorants de l'INRIA Rocquencourt, et a animé pour la troisième année consécutive le Colloquium Junior de l'Inria Rocquencourt. Il était intervenant à l'atelier « Physique nucléaire, mathématiques, modélisation » du Salon de l'Éducation (22 novembre 2000) et dans un débat sur la recherche en Europe au Palais de la Découverte qui a eu lieu le 18 octobre 2001, dans le cadre de la Fête de la Science. Il a coorganisé une rencontre entre doctorants et PME à l'École polytechnique qui a eu lieu en avril

2001. Il a soutenu sa thèse [2] le 25 juin 2001 et commencé un postdoc en octobre 2001, dans le groupe « Algorithms and Complexity » dirigé par Kurt Mehlhorn, au Max-Planck-Institut für Informatik à Sarrebruck (Allemagne).

Philippe Flajolet a été membre du comité de programme du « *Seventh Seminar on Analysis of Algorithms* » oragnisé cette année par B. Vallée à Tatihou, et qui a rassemblé une soixantaine de 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 cinquantaine de probabilistes, de combinatoriciens et algorithmiciens du 26 au 30 mars 2001. 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 2002). Il est éditeur de la revue *Random Structures and Algorithmes* (Wiley) et « *Honarary Editor* » de *Theoretical Computer Science* (Elsevier). En 2001, il a été membre de jury ou rapporteur des thèses ou habilitations suivantes : C. Banderier (D, Paris VI), S. Boucheron (H, Paris-Sud), A. Micheli (D, Marne-la-Vallée), J. Petit (D, Barcelona). P. Flajolet est par aileurs 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. En 2001, P. Flajolet a été membre du comité d'évaluation de l'Institut Gaspard-Monge (Informatique, Marne-la-Vallée) et membre de la commission de recrutement de professeur à l'École Polytechnique Fédérale de Lausanne. 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. Il a été distingué en 2001 par l'ISI (connu notamment pour la publication du « Science Citation Index ») comme l'un des 300 chercheurs français toutes disciplines confondues (et moins d'une dizaine d'informaticiens) en tant qu'auteur de plus de cinq articles d'impact majeur au cours des deux décennies écoulées.

Mireille Régnier a été rapporteur de la thèse de Cédric Chauve (Université de Bordeaux) et membre du jury de Jean Jabbour-Hattab (Université de Versailles).

Philippe Robert a été le rapporteur des thèses de A. Ben-Tahar et A. Yaacoubi de l'Université Hassan II à Casablanca.

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 Médicis. Il est également membre du comité éditorial du *Journal of Symbolic Computation* et du comité de programme de la conférence ISSAC'02. 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 a participé aux jurys de thèse de Dominique Rossin (École polytechnique), Grégoire Lecerf (École polytechnique, thèse qu'il co-encadrait avec M. Giusti), Olivier Cormier (Université Rennes 1) et Anne Fredet (École polytechnique). 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

Cyril Banderier, a enseigné, pour sa troisième et dernière année de monitorat, à l'université de Paris-Nord (P13) en premier cycle universitaire et a encadré un projet de maîtrise.

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

Marianne Durand a assuré des vacations à l'École polytechnique en tronc commun d'informatique. Elle est enseignante à l'université de Versailles–Saint-Quentin, ou elle assure un cours et un TD d'informatique en première année de DEUG.

Philippe Flajolet enseigne avec B. Vallée un cours « Modèles Combinatoires » au sein du DEA Algorithmique de la région parisienne.

Ludovic Meunier, moniteur à l'Université de Jussieu (P6), enseigne en premier cycle universitaire des travaux pratiques d'informatique.

Mireille Régnier enseigne un cours d'« Algorithmique et combinatoire » au DEA de bioinformatique d'Évry.

Philippe Robert donne un cours de DEA intitulé « Processus stochastiques » au DEA Maths-Info de l'Université de Versailles St-Quentin (avec J.-M Fourneau).

Bruno Salvy a donné un cours de Calcul Formel-Maple aux agrégatifs de l'École Normale Supérieure de Paris.

9.3 Participation à des colloques, séminaires, invitations

Cyril Banderier a donné des exposés aux Journées Aléa (Marseille), à l'École Jeunes Chercheurs de Lyon et à la rencontre « Analysis of Algorithms » à Tatihou. Il a par ailleurs assisté aux Douzièmes Rencontres Arithmétiques (Caen), à la conférence Random Strutures and Algorithms (Poznan, Pologne) et, dans le cadre du projet Européen Alcom-FT, à l'école Algorithm Engineering et au First Annual Review Meeting and Workshop (à Rome, en septembre 2001). Il a été invité à donner un exposé au Séminaire de théorie analytique des nombres (Bordeaux), et un autre au Dipartimento di Sistemi e Informatica (université de Florence), dans le cadre d'une invitation d'une semaine.

Frédéric Chyzak a présenté ses résultats sur la manipulation de suites et fonctions spéciales par le calcul formel au RISC (Linz, Autriche) lors d'une visite du 18 au 24 avril, au LRI (université d'Orsay), à l'équipe Algorithmique du GREYC (université de Caen), ainsi qu'au groupe *Computational Mathematics* de l'université de Kassel (Allemagne), où l'exposé théorique a été complété d'une démonstration logicielle de son programme MGFUN.

Marianne Durand a donné un exposé aux journées ALEA 2001 à Marseille. Elle a assisté à l'École Jeunes Chercheurs « Algorithmique et Calcul Formel » de Lyon, au séminaire AofA à Tatihou (France) et à l'école d'*algorithm engineering* à Rome (Italie). En avril, elle a également passé 2 semaines à l'université HKUST à Hong-Kong.

Philippe Flajolet a été invité au Colloquium de Mathématiques de l'Université de Limoges (février 2001), au Séminaire d'Algorithmique de Marne-la-Vallée, au Séminaire de Statistiques de l'Université Paris V, ainsi qu'au Colloque ASTI 2001 (Cité des Sciences) où il a donné une conférence d'intérêt général intitulée « Informatique et Mathématiques », ce au sein d'une session organisée par le Pr. Kahn. Il a par ailleurs été invité en octobre 2001 à l'Université de Purdue où il a présenté une conférence au séminaire « Mathematics of Engineering Sciences ». P. Flajolet est par ailleurs conférencier invité au prochain *International Congress of Mathematicians*, Beijing 2002, où il présentera une conférence sur le thème « *Singular Combinatorics* ».

Ludovic Meunier a assisté à l'École de Jeunes Chercheurs « Algorithmique et Calcul Formel » en février, à Lyon et à la 29ème École de printemps d'informatique théorique « Arithmétique des ordinateurs » en mars, à Prapoutel-les-Sept-Laux. Il a participé à la conférence *Mathematical Knowledge Management'01*, qui a eu lieu à Linz en Autriche.

Mireille Régnier a participé à la conférence RECOMB'01 et au colloque MFRS'01 à Montréal, au Workshop de l'OTAN sur les « Heuristiques en Bioinformatiques » à San Miniato en Italie. Elle a présenté une communication aux journées du LANFOR et à AoFa'01. Elle a fait des séminaires au LIAFA et à l'EIMB (Moscou).

Philippe Robert a participé à la conférence INFORMS à New York (USA) du 25 au 27 juillet pour présenter des travaux sur les problèmes d'allocation de bande passante [37]. Il a donné une conférence sur le comportement du protocole TCP à la revue du projet européen ALCOM-FT à Rome du 13 au 14 septembre. Invité à un workshop du 13 novembre au 15 novembre sur les réseaux à l'Université de Göteborg, il a donné une conférence sur les algorithmes de contrôle de la congestion dans les réseaux. Il a participé à la conférence Globecom du 26 au 28 novembre à San Antonio, Texas pour présenter l'article [25] sur le comportement du protocole TCP. Il a donné un exposé sur les algorithmes de contrôle de la congestion dans les réseaux à l'École Normale Supérieure et à l'Université de Versailles. Il a donné un exposé sur l'utilisation de modèles probabilistes dans l'étude des réseaux de télécommunication aux agrégatifs de l'École Normale Supérieure.

Bruno Salvy a été conférencier invité au *Séminaire Lotharingien de combinatoire* à Bertinoro (Italie). Il a présenté ses travaux communs avec L. Meunier à la conférence *Mathematical Knowledge Management'01*, à Linz (Autriche).

Mathias Vandenbogaert a participé à l'École Jeunes Chercheurs « Algorithmique et Calcul Formel » à l'ENS, Lyon, aux journées ALEA'2001 à Marseille (Mars 2001) et à JOBIM '01 à Toulouse (mai 2001). Il a effectué un séjour de 15 jours à NII-Genetika (Moscou). Il a présenté

une communication à WABI '01 (Aarhus, août 2001). Il a présenté son travail au LaBRI (Bordeaux), au LRI (Orsay) et au Colloque Junior de Rocquencourt. Il a assisté à aux colloques IMPG (Informatique, Mathématiques et Physique pour la Génomique) à Gif-sur-Yvette en mars 2001 et à Lyon en septembre 2001.

Bert Zwart à participé au workshop « The Mathematics of Stochastic Networks » à EURANDOM (Eindhoven) du 29 octobre au 2 novembre pour présenter la conférence « Fluid queues with M/G/infinity input ».

10 Bibliographie

Livres et monographies

[1] F. Chyzak (éditeur), *Algorithms Seminar, 2000–2001, Research Report*. Institut National de Recherche en Informatique et en Automatique, 2001. En préparation. 170 pages.

Thèses et habilitations à diriger des recherches

[2] C. Banderier, *Combinatoire analytique des chemins et des cartes*, thèse de doctorat, Université Paris VI, juin 2001.

Articles et chapitres de livre

[3] M. J. Atallah, F. Chyzak, P. Dumas, « A randomized algorithm for approximate string matching », *Algorithmica 29*, 3, 2001, p. 468–486.

[4] C. Banderier, P. Flajolet, G. Schaeffer, M. Soria, « Random Maps, Coalescing Saddles, Singularity Analysis, and Airy Phenomena », *Random Structures & Algorithms 19*, 3/4, 2001, 53 pages. In press.

[5] C. Banderier, P. Flajolet, « Basic Analytic Combinatorics of Directed Lattice Paths », *Theoretical Computer Science*, juillet 2001, To appear, 37 pages.

[6] P. Chassaing, P. Flajolet, « Hachage, Arbres, Chemins, et Graphes », *Bulletin de Probabilités 5*, 2001, p. 1–17.

[7] F. Chyzak, P. Paule, O. Scherzer, A. Schoisswohl, B. Zimmermann, « The construction of orthonormal wavelets using symbolic methods and a matrix analytical approach for wavelets on the interval », *Experimental Mathematics 10*, 1, 2001, p. 67–86.

[8] J. Clément, P. Flajolet, B. Vallée, « Dynamical Sources in Information Theory : A General Analysis of Trie Structures », *Algorithmica 29*, 1/2, 2001, p. 307–369.

[9] R. Cori, D. Rossin, B. Salvy, « Polynomial Ideals for Sandpiles and their Gröbner Bases », *Theoretical Computer Science*, 2001, à paraître. Version préliminaire dans le rapport de recherche INRIA n^0 3946, `http://www.inria.fr/rrrt/rr-3946.html`.

[10] J.-F. Dantzer, I. Mitrani, P. Robert, « Large Scale and Heavy Traffic Asymptotics for Systems with Unreliable Servers », *Queueing Systems, Theory and Applications 38*, 1, 2001, p. 5–24.

[11] V. Dumas, P. Robert, « On the throughput of a resource sharing model », *Mathematics of Operation Research 26*, 1, 2001, p. 163–173.

[12] M. Durand, S. Taylor, « Emerging Behavior as Binary Search Trees Are Symetrically Updated », *Theoretical Computer Science*, to appear, 17 pages.

[13] P. Flajolet, X. Gourdon, D. Panario, « The Complete Analysis of a Polynomial Factorization Algorithm over Finite Fields », *Journal of Algorithms 40*, 1, 2001, p. 37–81.

[14] P. Flajolet, K. Hatzis, S. Nikoletseas, P. Spirakis, « On the Robustness of Interconnections in Random Graphs : A Symbolic Approach », *Theoretical Computer Science*, 2001, In press. A preliminary version appears as INRIA Research Report **4069**, November 2000, 21 pages., `http://www.inria.fr/rrrt/rr-4069`.

[15] P. Flajolet, G. Louchard, « Analytic Variations on the Airy Distribution », *Algorithmica 31*, 3, 2001, p. 361–377.

[16] P. Flajolet, « $D \cdot E \cdot K = (100)_8$ », *Random Structures & Algorithms 19*, 3/4, 2001, Introduction to special volume on "Analysis of Algorithms" dedicated to D. E. Knuth ; 13pp. In press.

[17] M. Giusti, G. Lecerf, B. Salvy, « A Gröbner Free Alternative for Polynomial System Solving », *Journal of Complexity 17*, 1, mars 2001, p. 154–211.

[18] D. Hirschberg, M. Régnier, « Tight Bounds on the Number of String Subsequences », *Journal of Discrete Algorithms*, 2000, À paraître, version préliminaire présentée à CPM'99.

[19] P. Nicodème, B. Salvy, P. Flajolet, « Motif Statistics », *Theoretical Computer Science*, 2001, To appear. Extended version of an article published in the proceedings of 7th Annual European Symposium on Algorithms ESA'99, Prague, July 1999.

[20] D. Papatzenko, V. Makeev, A. Lifanov, M. Régnier, A. Nazina, C. Desplan, « Extraction of Functional Binding Sites from unique regulatory regions : The *Drosophila* early developmental enhancers », *Genome Research*, 2001, To appear ; preliminary version in Drosophila Workshop, Washington 2001.

[21] F. Tahi, M. Gouy, M. Régnier, « Automatic RNA secondary structure prediction with a comparative approach », *Computers and Chemistry*, 2001, poster at RECOMB'01 and MFRS'01 ; Montrèal ; To appear.

[22] B. Vallée, « Dynamical Sources in Information Theory : Fundamental Intervals and Word Prefixes », *Algorithmica 29*, 1/2, 2001, p. 262–306.

Communications à des congrès, colloques, etc.

[23] C. Banderier, « Factors' paradox », février 2001.

[24] A. Denise, M. Régnier, M. Vandenbogaert, « Assessing statistical significance of overrepresented oligonucleotides », *in : WABI'01*, Springer-Verlag, p. 85–97, 2001. in Proc. First Intern. Workshop on Algorithms in Bioinformatics, Aarhus, Denmark, August 20001 ; preliminary version as INRIA research report 4132.

[25] V. Dumas, F. Guillemin, P. Robert, « Limit results for Markovian models of TCP », *in : Globecom'2001*, San Antonio, novembre 2001.

[26] A. Dupuis, F. Guillemin, P. Robert, « Modeling Max-Min Fairness for Elastic Flows in Telecommunication Networks », *in : ITC'17*, Salvador da Bahia, décembre 2001.

[27] P. Flajolet, Y. Guivarc'h, W. Szpankowski, B. Vallée, « Hidden Pattern Statistics », *in : Automata, Languages, and Programming*, F. Orejas, P. Spirakis, J. van Leeuwen (éditeurs), *Lecture Notes in Computer Science*, 2076, Springer Verlag, p. 152–165, 2001. Proceedings of the 28th ICALP Conference, Crete, July 2001.

[28] L. Meunier, B. Salvy, « Automatically Generated Encyclopedia of Special Functions », *in : Mathematical Knowledge Management*, 2001. Proceedings MKM'01, Linz, September 2001.

[29] M. Mishna, « Attribute grammars and automatic complexity analysis », *in : FPSAC'01*, p. 371–380, mai 2001. Formal Power Series and Algebraic Combinatorics, Tempe, Arizona. An earlier version was INRIA Research Report 4021.

[30] M. Régnier, « Complexity of Unusual Words Counting », *in : JOBIM'00*, *Lecture Notes in Computer Science*, Springer-Verlag, 2000. à paraître, Proc. of JOBIM'00, Montpellier.

[31] M. Régnier, « Complexity of Unusual Words Counting », *in : JOBIM'00*, *Lecture Notes in Computer Science*, *2066*, Springer-Verlag, p. 101–117, 2001. Proceedings of JOBIM'00, Montpellier.

Rapports de recherche et publications internes

[32] C. Banderier, M. Bousquet-Mélou, A. Denise, P. Flajolet, D. Gardy, D. Gouyou-Beauchamps, « Generating Functions of Generating Trees », *Technical Report n° ALCOM FT-TR-01-17*, Alcom-FT Project, février 2001, 26 pages. Accepted for publication in *Discrete Mathematics*.

[33] A. Bostan, B. Salvy, É. Schost, « Fast Algorithms for Zero-Dimensional Polynomial Systems Using Duality », *Research Report n°4291*, Institut National de Recherche en Informatique et en Automatique, octobre 2001, 24 pages, `http://www.inria.fr/rrrt/rr-4291.html`.

[34] P. Flajolet, R. Sedgewick, « Analytic Combinatorics : Functional Equations, Rational and Algebraic Functions », *Research Report n°4103*, INRIA, 2001, 98 pages, `http://www.inria.fr/rrrt/rr-4103.html`.

[35] B. Salvy, « Even-Odd Set Partitions, Saddle-Point Method and Wyman Admissibility », *Research Report n°4201*, Institut National de Recherche en Informatique et en Automatique, 2001, 14 pages, `http://www.inria.fr/rrrt/rr-4201.html`.

Divers

[36] J.-F. Dantzer, V. Dumas, « Stability analysis of the Cambridge ring », 2001, Soumis.

[37] J.-F. Dantzer, P. Robert, « Analysis of a multi-class queueing system », 2001, Soumis.

[38] P. Duchon, P. Flajolet, G. Louchard, G. Schaeffer, « Random Sampling from Boltzmann Principles », Preprint, 2001, Submitted to ACM–STOC'2002, 12 pages.

[39] V. Dumas, F. Guillemin, P. Robert, « Effective bandwidths in a multiclass priority queueing system », *in : Infocom'2002*, 2001, Soumis.

[40] V. Dumas, F. Guillemin, P. Robert, « A Markovian analysis of AIMD algorithms », 2001, Soumis.

[41] M. Durand, « Asymptotic analysis of an optimized quicksort algorithm », Preprint, submitted to *Information Processing Letters.*, 2001, 8 pages.

[42] C. Fricker, P. Robert, D. Tibi, « On the fluid limits of some loss networks. », 2001, Soumis.

[43] M. Régnier, A. Denise, « Word Statistics conditioned by Overrepresented Words », 2001, en preparation.

[44] M. Régnier, A. Lifanov, V. Makeev, « Three Variations on Word Counting », submitted to BioInformatics ; preliminary version at GCB'00.