INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Action ADAGE

Algorithmique Discrète et ses Applications à la GÉnomique

Lorraine

THÈME 2B

Table des matières

ADAGE *est un avant-projet du LORIA (UMR 7503) commun au CNRS, à l'INRIA, à l'Université* HENRI POINCARÉ *Nancy 1, à l'Université Nancy 2 et à l'Institut National Polytechnique de Lorraine.*

1 Composition de l'équipe

Responsable scientifique
Grégory Kucherov [CR INRIA]
Assistant(e) de projet
Geneviève Grisvard-Pierrelée [AGT INRIA, jusqu'au 15/03/2001]
Franck Girault [du 17/04/2001 au 17/10/2001]
Hélène Zganic [TR INRIA, à partir du 01/09/2001, à 1/6 du temps]
Personnel CNRS
Jean-Luc Rémy[1] [CR]
Gilles Schaeffer [CR]
Personnel Université
Isabelle Debled-Rennesson [Maître de conférences IUFM de Lorraine]
Jocelyne Rouyer [Maître de conférences, UHP]
Poste d'accueil de spécialistes
Roman Kolpakov [INRIA, jusqu'au 31/08/2001]
Ingénieur associé
Ghizlane Bana [INRIA, à partir du 1/10/2001]
Chercheur invité
Ania Gambin [1 mois]
Stagiaires
Emmanuelle Becker [ENS Lyon, 1,5 mois]
Patricia Lavigne [DESS bioinformatique de Rouen, à partir du 29/10/2001]
Ralph Rabbat [MIT, 3 mois]
Alexei Stanger [Paris 6, 5 mois]

2 Présentation et objectifs généraux

L'avant-projet ADAGE a été créé au 1/1/2001 suite à la restructuration du projet POLKA. L'objectif général d'ADAGE consiste à mettre au point des algorithmes efficaces sur les structures discrètes (telles que mots, arbres, graphes, cartes, polyominos, ...). Cet objectif nous conduit à étudier en profondeur des propriétés combinatoires de ces structures, qui peuvent être de nature exacte ou probabiliste.

Nos recherches sont structurées en trois actions. La première porte sur l'algorithmique et la combinatoire des mots. Ici, nous travaillons sur l'analyse de complexité de problèmes sur les mots (textes, ou séquences de caractères) et sur le développement d'algorithmes efficaces d'analyse de mots. La deuxième action de recherche relève du domaine de la géometrie discrète. Les structures étudiées ici sont des objets géométriques discrétisés, décrits par un ensemble de

[1] affecté à partir du 14 mai 2001 aux activités syndicales à hauteur de 492 heures annuelles

points dans $\mathbb{Z}^2$ ou $\mathbb{Z}^3$. Comme dans le cas précédent, nous cherchons à mettre au point des algorithmes efficaces sur ces structures, qui vérifient leurs propriétés ou calculent des paramètres géométriques. La troisième action considère les modèles discrets sous un angle probabiliste, en supposant en général une distribution de probabilité sur l'espace de modèles possibles. Nous nous intéressons donc aux propriétés typiques des structures en question, et nous nous attachons en particulier aux méthodes d'analyse de ces propriétés.

Le champ d'application privilégié de nos travaux est la bioinformatique, domaine dans lequel les modèles discrets apparaissent de façon naturelle et essentielle. Ici, nous poursuivons des collaborations actives avec des équipes de biologistes sur des problèmes d'analyse de séquences d'ADN et de protéines.

Nous prêtons une attention particulière à la mise en place de logiciels expérimentaux basés sur des algorithmes issus de nos travaux. Deux logiciels d'analyse de texte sont développés dans l'avant-projet : le premier, grappe, permet de rechercher très rapidement dans un texte des motifs de certains types (sous-classes d'expressions régulières) ; le deuxième, appelé mreps, recherche des fragments répétés et avec la possibilité de variation entre les copies répétées. Les deux logiciels ont une version spécialisée pour le traitement de séquences d'ADN.

3 Fondements scientifiques

Mots clés : algorithmique discrète, structures discrètes, complexité, algorithmique des mots, recherche de motifs, géométrie discrète, aléa discret, analyse d'algorithmes.

Si l'on voulait définir la problématique de notre avant-projet par approximations successives, il serait naturel de commencer par la placer dans le domaine de l'*algorithmique discrète*. Construire un modèle discret d'un problème ou d'un phénomène du monde réel fait appel, sur le plan mathématique, aux *structures discrètes*, telles que graphes, mots, arbres, ensembles de points dans un espace, etc.

Pour pouvoir utiliser les modèles discrets, nous sommes donc amenés à étudier les propriétés des structures impliquées. En tant qu'informaticiens, nous nous intéressons aux *propriétés algorithmiques*, en particulier à l'*efficacité* (*complexité*) des calculs impliqués, que ce soit *en moyenne* ou *dans le cas le pire*.

Pour pouvoir développer des algorithmes efficaces sur les structures discrètes, ainsi que pour analyser et optimiser ces algorithmes, il faut donc comprendre et maîtriser les propriétés des structures sous-jacentes. Ces propriétés peuvent être de natures différentes : s'il s'agit de propriétés de nature exacte, on parle de *propriétés combinatoires* ; si le modèle en question est défini en termes probabilistes, c'est-à-dire via une distribution de probabilité dans un univers de modèles possibles, on aura affaire à des *propriétés typiques* (ou *statistiques*).
Nous allons maintenant brièvement présenter le domaine scientifique de chacune de nos actions de recherche.

Algorithmique des mots L'algorithmique des mots (ou l'algorithmique des séquences) est un domaine qui a vu un progrès considérable ces dernières années, comme le témoigne

la parution récente de quelques monographies sur ce sujet [CR94,Ste94,Gus97]. Tout en étant une partie intégrante de l'algorithmique discrète en général, l'algorithmique des mots forme aujourd'hui un domaine de recherche en soi, de même que l'algorithmique des graphes par exemple. Les progrès dans ce domaine ont été largement alimentés par ses nombreux champs d'application, dont deux – la bioinformatique et la recherche d'information sur l'Internet – sont particulièrement d'actualité aujourd'hui.

Les algorithmes sur les mots ont également un grand intérêt du point de vue théorique. À la base de cette théorie se trouvent quelques algorithmes et structures de données qui font partie du « trésor de l'algorithmique ». Le plus connu est probablement l'algorithme de Knuth-Morris-Pratt que l'on trouve dans tous les manuels d'enseignement d'algorithmique, mais qui, par ailleurs, a eu beaucoup d'applications à des problèmes divers (dont on continue à découvrir des exemples) et qui a fait l'objet d'une analyse mathématique intéressante et non-triviale [KMP77]. D'autres algorithmes textuels jouent également un rôle fondamental, comme l'algorithme de recherche par programmation dynamique d'une plus longue sous-séquence commune à deux séquences, dont les applications sont diverses et variées (comme l'utilitaire `diff` d'UNIX par exemple ou encore l'algorithme bien connu de Smith et Waterman d'*alignement local* de séquences biologiques). Parmi d'autres algorithmes, probablement moins connus mais tout aussi élégants, notons l'algorithme de recherche en temps linéaire de carrés dans un mot [Cro83] ou celui de recherche de palindromes, également en temps linéaire [Man75].

En outre, l'algorithmique des mots a développé des structures de données très puissantes, telles que l'arbre des suffixes (*suffix tree*) ou le DAWG (*Directed Acyclic Word Graph*). Le premier objectif de ces structures est de servir d'outils d'*indexation* de textes, c'est-à-dire de fournir une représentation spéciale de textes permettant d'exécuter efficacement des requêtes diverses. De plus, la transformation d'un texte en cette représentation se fait aussi très efficacement, à savoir en ligne et en temps linéaire. Une fois cette représentation obtenue, de nombreuses tâches peuvent être accomplies très efficacement. Sans essayer de les énumérer, nous renvoyons au livre [Gus97] dont une grande partie est consacrée aux diverses utilisations de l'arbre des suffixes.

Un aspect très important pour nous de l'algorithmique des mots est qu'elle s'appuie d'une façon essentielle sur les propriétés combinatoires des mots. De nombreux algorithmes utilisent dans leur fonctionnement, ou dans leur analyse, des théorèmes de la combinatoire des mots. C'est pourquoi la combinatoire des mots tient une place importante dans nos études.

En résumé, notre objectif consiste à mettre au point de nouveaux algorithmes efficaces d'analyse de mots, en s'appuyant sur des propriétés combinatoires des mots. L'application directe de ces algorithmes est l'analyse de séquences biologiques dont nous parlerons dans la

[CR94] M. Crochemore, W. Rytter, *Text algorithms*, Oxford University Press, 1994.

[Ste94] G. Stephen, *String Searching Algorithms*, World Scientific, Singapore, 1994.

[Gus97] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997.

[KMP77] D. Knuth, J. Morris, V. Pratt, « Fast pattern matching in strings », *SIAM J. Comput. 6*, 1977, p. 323–350.

[Cro83] M. Crochemore, « Recherche linéaire d'un carré dans un mot », *Comptes Rendus Acad. Sci. Paris Sér. I Math. 296*, 1983, p. 781–784.

[Man75] G. Manacher, « A new linear-time on-line algorithm for finding the smallest initial palindrome of the string », *J. ACM 22*, 1975, p. 346–351.

section 4.1.

Géométrie discrète Parmi les structures discrètes que nous étudions figurent les ensembles discrets du plan ou de l'espace. La géométrie discrète, qui étudie ces objets, est apparue dans les années 70. Elle a pour objectif de définir un cadre théorique pour transposer dans $\mathbb{Z}^n$ les bases de la géométrie euclidienne, les notions discrètes définies étant le plus proche possible des notions continues que nous connaissons (telles que distance, longueur, convexité, ...). Plusieurs façons d'aborder cette étude ont été développées [CM91] :

- le point de vue topologique s'intéressant par exemple à l'équivalent discret du théoréme de Jordan (toute courbe fermée simple sépare le plan en deux domaines : l'intérieur et l'extérieur de la courbe),
- le point de vue morphologique qui étudie les transformations de formes,
- le point de vue arithmétique, introduit par Jean-Pierre Reveillès en 1989 [Rev91], qui donne une définition en compréhension des droites et plans discrets.

C'est cette dernière approche que nous utilisons. Une droite discrète du plan est ainsi l'ensemble des points de coordonnées (x, y) de $\mathbb{Z}^2$ vérifiant une double inégalité de la forme $\mu \leq ax + by < \mu + \omega$, avec a, b, μ, ω entiers. Les propriétés des droites discrètes définies de la sorte sont en relation étroite avec les propriétés des nombres entiers et nous rapprochent ainsi de la combinatoire des mots (utilisation des mots de Sturm par exemple). Les plans discrets sont définis de manière analogue.

Ces définitions analytiques permettent de représenter de manière compacte des objets discrets, d'étudier des objets intrinséquement discrets (pas uniquement des approximations d'objets continus), et de définir des objets discrets infinis.

De nombreux résultats fondés sur cette approche ont vu le jour ces dix dernières années :

- définition et étude de nouvelles classes d'objets discrets (droites 3D, hyperplans, cercles, sphères, simplexes, ...),
- reconnaissance analytique permettant, non seulement de dire si une suite de points est ou non un segment de droite, mais aussi de donner les coefficients des inéquations analytiques correspondantes,
- reconstruction analytique visant à passer d'une représentation en compréhension du discret à une représentation en compréhension du continu,
- transformations discrètes : applications quasi-affines, rotations, filtres,
- visualisation, utilisant des propriétés des objets discrets telles que l'épaisseur optimale des objets.

En ce qui nous concerne, nos travaux se rangent principalement parmi les trois premières thématiques de cette liste.

Aléa discret L'aléa discret est le champ d'étude consacré aux propriétés typiques de structures combinatoires aléatoires. Il est maintenant classique en informatique de considérer à côté des analyses de pire cas, des analyses en moyenne pour des modèles de données aléatoires ou

[CM91] J.-M. Chassery, A. Montanvert, *Géométrie discrète en imagerie*, Hermès, Paris, 1991.

[Rev91] J.-P. Reveillès, *Géométrie discrète, calculs en nombre entiers et algorithmique*, Thèse d'état, Université Louis Pasteur, Strasbourg, 1991.

pour des algorithmes probabilistes. Ce type d'analyses, popularisées par Knuth dans *The Art of Computer Programming*, se développe activement pour traiter des modèles de plus en plus réalistes, et donc plus complexes, en s'appuyant très largement sur le progrès des techniques de combinatoire énumérative (en particulier analytique, mais aussi algébrique et bijective) et des méthodes de génération aléatoire.

Du point de vue des outils mathématiques nous faisons appel et développons au premier chef des méthodes d'énumération. Pour étudier le comportement d'un paramètre combinatoire (moyenne, variance, distribution), notre approche s'appuie sur le codage global de l'information recherchée par des séries génératrices, accessibles au travers de décompositions combinatoires et d'équations fonctionnelles associées. Cette approche, initiée en analyse d'algorithmes par Knuth puis Flajolet, Odlyzko, Sedgewick et d'autres[FO90,FS01], permet de traiter de larges classes de problèmes qui correspondent à des classes d'équations et de séries génératrices bien comprises (telles que les classes rationnelles ou algébriques[Sta01]). La problématique évolue ainsi de l'automatisation de l'analyse dans les cas les plus favorables au traitement d'instances particulièrement pointues. Entre ces deux extrêmes se situe par exemple l'analyse de modèles combinatoires qui conduisent à un type particulier d'équations (dites aux variables catalytiques), dont plusieurs instances sont résolues sans que le statut de la classe entière soit encore bien clair.

Une spécificité de notre approche est à chercher dans notre intérêt particulier pour l'aspect expérimental, au travers de la génération aléatoire. Il est peut-être utile de rappeler ici que la génération aléatoire peut être utilisée d'une manière très similaire à l'analyse en moyenne, et la complète notamment lorsqu'on atteint les limites des techniques actuelles d'analyses en moyenne. L'utilisation d'un générateur aléatoire ne fournit certe que des résultats expérimentaux, mais permet d'observer des paramètres autrement inaccessibles et souvent de formuler des conjectures qui dirigent l'analyse.

Les méthodes de génération aléatoire se rangent en première analyse dans deux catégories : d'un côté les méthodes de marches aléatoires markoviennes[2], et de l'autre les méthodes combinatoires. Les premières réalisent une marche aléatoire dans l'espace des configurations possibles jusqu'à avoir oublié tout de leur point de départ. Les secondes au contraire cherchent à construire directement un objet aléatoire, en tirant partie d'informations structurelles [FZVC94]. Ces deux domaines sont en pleine expansion, en particulier suite aux avancées spectaculaires réalisées sur la maîtrise des temps de convergence pour les approches markoviennes, et à l'utilisation d'algorithmes probabilistes dans les approches combinatoires. C'est dans cette seconde tendance que se situent nos travaux.

[2] cf. `http://dbwilson.com/exact/`

[FO90] P. Flajolet, A. Odlyzko, « Singularity analysis of generating functions. », *SIAM J. Discrete Math. 3*, 2, 1990, p. 216–240.

[FS01] P. Flajolet, R. Sedgewick, *The average case analysis of algorithms*, 2001, Livre en préparation, certaines parties disponibles comme rapports INRIA.

[Sta01] R. P. Stanley, *Enumerative combinatorics. Volume 2. Paperback ed.*, Cambridge Studies in Advanced Mathematics. 62. Cambridge: Cambridge University Press. xii, 585 p. , 2001.

[FZVC94] P. Flajolet, P. Zimmerman, B. Van Cutsem, « A Calculus for the Random Generation of Labelled Combinatorial Structures », *Theoretical Computer Science 132*, 1-2, 1994, p. 1–35.

4 Domaines d'applications

4.1 Bioinformatique

Mots clés : biologie, bioinformatique, séquence d'ADN, séquence de protéine, gène, promoteur, alignement de séquences.

Les modèles discrets apparaissent dans tous les domaines d'applications, mais il en est un qui joue pour nous un rôle tout à fait particulier. Il nous sert d'une part de source de problèmes et d'autre part de domaine privilégié pour tester et appliquer nos idées et méthodes. Il s'agit de la biologie moléculaire, c'est-à-dire de l'étude de macromolécules biologiques (ADN, ARN, protéines). L'irruption de modèles discrets dans ce domaine est due à la découverte de la structure de ces molécules, laquelle s'est avérée être un enchaînement linéaire d'éléments constituants qui appartiennent à un petit nombre de types. Ceci justifie immédiatement l'adoption d'un modèle discret de ces molécules : dans leur forme linéaire, ces molécules sont représentées par des chaînes de lettres tirées d'un alphabet de petite taille. Bien que ce modèle linéaire ne capture pas, du moins d'une façon adéquate, toutes les propriétés des molécules biologiques, il en capture une grande partie, et nos études portent en général sur des propriétés biologiques reflétées au niveau des séquences. Autrement dit, nous nous intéressons aux « empreintes » de phénomènes biologiques dans les séquences nucléiques ou protéiques. Ces « empreintes » sont décrites à l'aide de motifs, et une de nos motivations consiste à faire profiter la bioinformatique des techniques d'analyse et de recherche de motifs développées en algorithmique et en analyse probabiliste.

Ci-dessous sont présentées trois actions de recherche que nous poursuivons en collaboration avec des biologistes. Les deux premières (sections 4.1.1 et 4.1.2) possèdent, entre autre, une caractéristique commune : elles font intervenir une notion de *significativité* ou de *pertinence* d'un événement observé. Cette notion apparaît dans la situation suivante, très générale en bioinformatique.

Au cours du traitement informatique, des événements sont détectés (par exemple, des événements au niveau des séquences, tels que similarités ou répétitions), dont la véritable valeur biologique ne peut être déterminée automatiquement sans intervention des biologistes. Cependant au vu de la quantité de données à traiter, il est nécessaire d'avoir un critère pour éliminer *a priori* un maximum d'événements. Pour cela on s'intéresse à la vraisemblance de l'événement sous l'hypothèse où cet événement n'aurait aucune cause biologique, mais serait uniquement dû au hasard. Cette hypothèse est appelée l'*hypothèse nulle* et largement utilisée en biologie pour effectuer un prétraitement des données. L'idée sous-jacente de cette approche est que si l'occurrence d'un événement peut s'expliquer par le hasard, il est peu probable qu'il reflète un mécanisme biologique. *A contrario*, seuls des événements « surprenants » peuvent être biologiquement significatifs. Un exemple devenu classique[MMP01] est l'abondance « anormale » dans le génome de *E.coli* du motif `gctggtgg` dont la fonction biologique a été mise en évidence.

Pour chaque type d'expérience, la mise en pratique de cette approche nécessite un modèle probabiliste du phénomène étudié et le calcul des probabilités correspondantes. L'exemple le

[MMP01] F. Muri-Majoube, B. Prum, « Une approche statistique de l'analyse des génomes », *in : Gazette des mathématiciens, 89*, 2001.

plus connu est sans doute le modèle de Karlin utilisé par `blast`, logiciel de loin le plus utilisé en bioinformatique, qui recherche des similarités locales entre une protéine dite *query* et les séquences d'une base de données. Cependant dans de nombreux cas, il n'y a pas encore de modèle mathématique satisfaisant et on se contente d'utiliser des mesures empiriques de la qualité des résultats trouvés qui, à la différence des arguments statistiques, posent de gros problèmes de normalisation. De tels scores seront typiquement capables de classer les résultats d'une expérience donnée du plus au moins intéressant, mais ne permettent pas de comparer des expériences différentes, comme le permet par exemple la fonction *expect* de `blast`. Par conséquent, il est important de définir un modèle probabiliste et de pouvoir calculer la notion de pertinence sous-jacente.

Du point de vue méthodologique, de nombreux travaux font appel à des modélisations probabilistes asymptotiques (*e.g.* le modèle de Karlin). Cependant, les conditions pratiques d'application sont très fréquemment largement en dessous des régimes asymptotiques et les effets de taille finie doivent être pris en compte. Au contraire, les modélisations issues des méthodes de l'aléa discret (voir la section 6.3) travaillent directement dans le régime fini, ce qui donne des modèles plus adaptés. C'est ce qui rattache ces questions aux problématiques fondamentales de notre avant-projet.

Un autre lien entre nos actions bioinformatiques et nos recherches fondamentales passe naturellement par l'algorithmique des mots (voir la section 6.1). Cet aspect est présent dans toutes nos actions, et surtout dans l'action 4.1.3 qui est un « pur produit » de nos travaux théoriques en algorithmique. Notons que le modèle de répétitions, sous-jacent dans l'action 4.1.3, ne fait pas appel à un modèle probabiliste ; il est basé sur une définition purement combinatoire et n'utilise pas la notion de pertinence. En revanche, l'action 4.1.1 comporte à la fois une composante combinatoire (extraction de motifs candidats) et probabiliste (estimation de la pertinence de ces motifs).

4.1.1 Analyse de promoteurs dans un génome bactérien

Dans le cadre du Thème « Bioinformatique et Applications à la Génomique » du PRST Intelligence Logicielle, nous travaillons sur l'identification et la classification des promoteurs dans un génome bactérien, en collaboration avec des chercheurs du Laboratoire de Génétique et de Microbiologie de l'Université Henri Poincaré de Nancy (Pierre Leblond, Bertrand Aigle). L'objectif de ce travail est l'identification des sites de fixation dans les zones promotrices du génome de la bactérie *Streptomyces cœlicolor*. Notons que cette bactérie a un intérêt économique certain, en particulier du fait que près de 70% des antibiotiques utilisés dans l'industrie pharmaceutique contiennent une substance produite par des bactéries de ce type.

Le problème consiste à identifier, dans les régions en amont des parties codantes du génome, les sites de fixation des facteurs σ. Ces derniers sont des protéines faisant partie de l'ARN polymérase – complexe moléculaire qui assure la transcription de gènes, première étape vers la production de protéines. Les facteurs σ sont responsables de la fixation de l'ARN polymérase sur la molécule d'ADN, ce qui initialise le processus de transcription et détermine, en particulier, le début de la séquence transcrite, appelé la position +1 ou le *TSS (Transcription Start Site)*.

Il est connu que dans les bactéries (organismes procaryotes), les endroits de séquences

reconnus par un facteur σ sont localisés aux positions -10 et -35 par rapport au *TSS*. Cela nous conduit à rechercher des motifs composés de deux *boîtes*, séparées par une distance de 25 nucléotides environ.

Une première étape du travail a consisté à observer le phénomène dans 150 séquences étudiées biologiquement pour lesquelles les *TSS* sont connus ainsi que quelques motifs de sites de fixation. Ces résultats figurent dans deux publications [BB95,Str92]. Ces 150 séquences furent utilisées pour tester les différentes approches envisagées. Pour élaborer la stratégie de recherche des motifs des sites de fixation, nous sommes partis de l'idée qu'un motif de site de fixation devait être « surprenant » par rapport à son environnement, c'est-à-dire tel que le nombre d'apparitions dans les séquences soit anormalement grand par rapport à celui prévu par « le hasard ». Nous avons d'abord identifié, à l'aide d'un algorithme purement combinatoire, les motifs qui vérifient un certain seuil d'apparition et sont donc les premiers candidats pour correspondre aux sites de fixation. Pour déterminer le « taux de surprise » de ces motifs, nous avons utilisé le logiciel R'MES[3] qui identifie les motifs dont la fréquence d'apparition dans une séquence d'ADN est inattendue. Ce logiciel utilise des modèles de chaînes de Markov d'ordre m et calcule le score d'un motif en fonction de sa fréquence attendue et de sa fréquence observée. Les résultats obtenus par ce logiciel ont été exploités et traités ; pour chaque motif ayant obtenu un score remarquable, son alignement potentiel à une distance -35 ou -10 du point de départ de la transcription d'un gène a été testé. Cette approche s'est révélée positive pour l'étude des 150 séquences ; en effet, certains motifs des sites de fixation identifiés dans les publications d'une façon expérimentale ont été clairement reconnus. En revanche, certains motifs présentés dans les publications comme pouvant correspondre à des sites de fixation ont été réfutés par notre méthode.

La seconde étape a été de tester cette méthode sur le génome entier de *Streptomyces cœlicolor* contenant environ 7000 gènes. Notons qu'ici nous ne connaissions pas les positions des *TSS* et par conséquent nous ne pouvions donc pas tester si les motifs trouvés se situent dans les régions -35 ou -10. L'objectif est donc de retrouver les sites « à l'aveugle », sans pouvoir les valider à l'aide d'alignements. Les premiers résultats obtenus sont très bruités. D'autres critères biologiques doivent donc être considérés, l'analyse de significativité seule n'étant probablement pas suffisante pour discriminer les sites. Nous poursuivons ce travail avec l'objectif d'inclure dans l'analyse des critères biologiques non encore testés, tels que, par exemple, le niveau d'apparition de motifs identifiés dans les parties codantes, ou d'autres critères pouvant être mis en place en collaboration avec les biologistes.

4.1.2 Calcul de score pour l'alignement de séquences protéiques

Le problème de l'alignement des séquences de protéines et, partant, de la recherche de motif dans ces séquences est un domaine de compétence reconnu de l'équipe de bioinformatique de

[3] `http://www.inra.fr/bia/J/AB/genome/RMES2`

[BB95] W. R. Bourn, B. Babb, « Computer assisted identification and classification of streptomycete promoters », *Nucleic Acids Research 23*, 18, 1995, p. 3696–3703.

[Str92] W. R. Strohl, « Compilation and analysis of DNA sequences associated with apparent streptomycete promoters », *Nucleic Acids Research 20*, 5, 1992, p. 961–974.

l'IGBMC à Strasbourg, avec laquelle nous développons une collaboration dans le cadre du génopôle Strasbourg-Alsace-Lorraine.

D'un point de vue informatique, les problèmes de recherche de motifs et d'alignement de séquences de protéines semblent identiques aux problèmes correspondants pour les séquences nucléotidiques : seule change *a priori* la taille de l'alphabet. Pourtant au contact des biologistes on découvre qu'il n'en est rien et que les problématiques sont très différentes : les taux de similarité entre objets à comparer sont plus faibles et entrent en jeu les propriétés chimiques des acides aminés. Tout ceci conduit à des problèmes d'optimisation à critères multiples qui rendent les définitions de scores parfois assez arbitraires.

Dans ce cadre, nous nous concentrons sur des problèmes de score directement rencontrés dans les logiciels développés par l'équipe de l'IGBMC, en particulier sur l'amélioration du score utilisé par le logiciel `Ballast`[PTP00]. Il s'agit de passer pour ces applications précises d'une approche de type score, à une approche plus fondée statistiquement (voir la discussion au début de la section 4.1).

Ce logiciel traite un problème d'alignement local de séquences de protéines. Il retraite la sortie du logiciel standard `blast` qui recherche dans une base de données les séquences présentant une similarité avec une séquence *query*. Alors que `blast` utilise uniquement les qualités propres de chaque similarité pour calculer son score, `Ballast` se concentre sur des segments privilégiés (zones de la *query* rencontrant de nombreuses similarités dans la base) et néglige les similarités isolées, jugées peu importantes. Le score de `Ballast` est ainsi plus significatif que celui de `blast` mais, en l'absence de traitement statistique, n'est pas normalisé.

Nous collaborons donc à une nouvelle version de ce logiciel, au sein de laquelle la méthode de score est modifiée de façon à permettre une évaluation statistique approchée, sous forme de *p-values* associées aux scores. Ceci nous a conduit, avec Frédéric Plewniak et Olivier Poch de l'IGBMC, à redéfinir plusieurs étapes du traitement et en particulier à traiter algorithmiquement et statistiquement le problème de la fragmentation des segments privilégiés et de leur redondance. Le nouvel algorithme ainsi défini est en cours d'implantation et de test au sein de l'IGBMC.

4.1.3 Recherche de répétitions dans les séquences d'ADN

La séquence d'un génome peut être vue comme une combinaison de motifs qui s'entrelacent pour former une mosaïque extrêmement complexe. Mis à part reconnaître des motifs tels que les séquences promotrices (voir la section 4.1.1), il est intéressant de mettre en évidence les répétitions de certains mots. En effet, même s'il paraît plus « raisonnable » pour un génome de ne posséder aucune redondance, dans la réalité ce n'est pas ce que l'on observe et dans certains cas les répétitions peuvent être fortement sélectionnées au cours de l'évolution. Plus particulièrement, les répétitions en tandem (répétitions successives) qui ont la particularité de varier en nombre de copies (répétitions polymorphes), peuvent être impliquées dans des maladies génétiques humaines comme le diabète insulino-dépendant. Dans cette maladie, c'est une variation dans le nombre de répétitions d'une séquence en tandem, située à quelque distance du gène de l'insuline, qui est responsable de la susceptibilité à la maladie.

[PTP00] F. Plewniak, J. Thompson, O. Poch, « Ballast: Blast post-processing based on locally conserved segments », *Bioinformatics 16*, 2000, p. 750–759.

Chez la bactérie, l'instabilité de répétitions en tandem permet de modifier l'expression des gènes intervenant directement dans la virulence bactérienne. En effet, les répétitions en tandem se trouvent associées aux gènes en faisant partie de leurs régions codantes ou bien de leurs régions régulatrices. L'étude des répétitions peut nous renseigner également sur d'autres phénomènes cellulaires et évolutifs comme chez certaines bactéries où la présence de répétitions est probablement liée à l'existence d'éléments transférés horizontalement (les transposons). Par ce mécanisme, les bactéries acquièrent de nouvelles fonctions. De telles répétitions peuvent par ailleurs avoir un intérêt pratique comme l'établissement de profils génétiques. Il suffit d'identifier les répétitions en tandem et de s'assurer de leurs polymorphismes dans l'espèce considérée.

Il est donc très important de pouvoir rechercher les répétitions de façon rapide et exhaustive, et le logiciel `mreps` décrit dans la section 5.2 fournit un moyen puissant pour cette tâche. Cet outil pourrait venir en aide aux biologistes dans l'analyse exploratoire des génomes mais aussi dans leurs travaux de cartographie des génomes bactériens déjà séquencés. L'efficacité de mreps est telle que l'on peut lui soumettre une séquence de plusieurs dizaines de mégabases mais surtout qu'il repère des motifs ayant des périodes (la taille de l'élément répété) importantes, allant jusqu'à plusieurs dizaines de kilobases, que d'autres outils ne mettent pas en évidence. Nous avons appliqué mreps à plusieurs organismes mais notre intérêt s'est porté sur la bactérie *Neisseria meningitidis* MC58, qui est un agent bactérien de la méningite dont le génome fait 2,27 mégabases. L'application de mreps a permis de révéler la répétition exacte d'une région codante de 32 kilobases environ. En parallèle avec les tests de mreps sur plusieurs génomes et l'analyse de ces résultats, nous cherchons actuellement à identifier de nouveaux cas de figure non répertoriés dans la littérature.

5 Logiciels

5.1 grappe

Mots clés : analyse de texte, séquences d'ADN, recherche de motifs, motif multiple, motif avec espace.

grappe est un logiciel qui recherche dans un texte plusieurs motifs simultanément, chacun étant composé d'une suite de fragments (mots) séparés par des espaces de longueur *a priori* non-bornée. Ce logiciel, déposé à l'APP l'année dernière, est diffusé par plusieurs voies :

- à partir de la page des logiciels développés à l'INRIA `http://www.inria.fr/valorisation/logiciels/index.fr.html`, ainsi que sur le CD ROM des logiciels libres de l'INRIA,
- à l'adresse `http://www.loria.fr/~kucherov/SOFTWARE/grappe-3.0/grappe-3.0-en.html`,
- à partir de cette année, depuis la page de la plateforme *Qualité et sûreté des logiciels* `http://qsl.loria.fr/` dont grappe fait partie.

Notons qu'il existe une version spécialisée de grappe pour le traitement de séquences d'ADN/ARN. Notons en outre que nous utilisons grappe dans le travail sur l'analyse de promoteurs, décrit dans la section 4.1.1.

5.2 mreps

Mots clés : séquence d'ADN, recherche de répétitions, répétition maximale, répétition en tandem.

mreps est un logiciel de recherche de répétitions dites maximales dans un fichier texte en général et dans une séquence d'ADN en particulier. Les répétitions maximales sont des répétitions successives, appelées parfois *périodicités* dans la littérature informatique et *répétitions en tandem* dans la littérature génomique. La naissance de mreps, il y a deux ans, a suivi les travaux théoriques [KK99] dans lesquels nous avons proposé un algorithme très efficace (en temps linéaire) pour rechercher toutes les répétitions maximales exactes dans un texte.

Depuis, nous poursuivons le développement de ce logiciel à la fois sur le plan théorique (voir la section 6.1) et appliqué. Cette année, ce développement a vu un progrès très considérable. Deux avancées majeures ont eu lieu. Le module de factorisation a été reprogrammé, parce qu'il constituait un « goulot d'étranglement » du logiciel quant à la taille de mémoire requise. Cela a permis de passer à une échelle de taille beaucoup plus grande, à savoir de traiter des séquences entières de dizaines de millions de caractères, sans utiliser de fenêtre glissante. Une autre amélioration importante a consisté en l'intégration d'un module de recherche de répétitions approchées, selon l'algorithme que nous avons récemment mis au point (voir la section 6.1). En plus de ces deux modifications, d'autres changements ont été apportés au logiciel dans le but d'étendre ses fonctionnalités ou d'améliorer la convivialité, tels que la possibilité de sortie au format XML ou encore l'ajout de nombreux paramètres permettant de spécifier le type des répétitions recherchées.

Ces développements ont donné lieu à une nouvelle version de mreps, spécialisée pour le traitement de séquences d'ADN. La possibilité de rechercher, de façon instantanée, dans des génomes entiers (rappelons que la taille typique d'un génome bactérien est de l'ordre de quelques millions de caractères) *toutes* les répétitions en tandem avec un taux d'erreurs donné, fait en effet de mreps un logiciel très intéressant pour les bioinformaticiens. Il n'a pas de concurrent en France, et son seul concurrent sur le plan international, créé à la *Mount Sinai School of Medicine* à New-York, est fondé sur un algorithme probabiliste et possède une limitation forte sur la période des répétitions (500 dans la version diffusée actuellement). Or, il existe des répétitions dans les séquences d'ADN dont l'élément répété a une taille bien plus grande (voir la section 4.1.3).

La version 2.0 de mreps incluant ces nouvelles fonctionnalités est diffusée à l'adresse `http://www.loria.fr/~kucherov/SOFTWARE/mreps/`. Une interface Web de mreps est également accessible à cette adresse. Enfin, d'autres améliorations de mreps sont envisagées.

[KK99] R. Kolpakov, G. Kucherov, « Finding Maximal Repetitions in a Word in Linear Time », *in : 1999 Symposium on Foundations of Computer Science , FOCS'99, New-York , USA*, IEEE Computer Society, IEEE Computer Society, p. 596–604, 1999.

6 Résultats nouveaux

6.1 Algorithmique des mots

Les travaux sur la recherche de *répétitions maximales* ont été poursuivis cette année. L'objectif général de ces travaux est de développer des algorithmes efficaces d'identification de répétitions dans les mots, en particulier de répétitions successives (périodicités). Dans ce domaine, que nous étudions déjà depuis quelques années, nous avons obtenu des résultats importants : dans nos travaux de 1998-1999, nous avons mis au point un algorithme de recherche de toutes les répétitons maximales exactes dans un mot en temps $O(n)$ (n étant la longueur du mot). Comme c'est souvent le cas dans l'algorithmique des mots, cet algorithme et la preuve de sa linéarité sont fondés sur des propriétés combinatoires profondes. L'année dernière, nous avons étendu cette technique à la recherche de *répétitions à espace fixe*, c'est-à-dire à la recherche de toutes les paires d'occurrences d'un même facteur séparées par une distance constante d spécifiée par avance. Pour ce problème, nous avons élaboré un algorithme en $O(n \log d + S)$, où S est le nombre de répétitions trouvées.

Une suite logique de cette recherche a consisté à admettre une certaine variation entre les copies d'un facteur répété. Autrement dit, il s'agit de l'identification de *répétitions approchées* et non plus de répétitions exactes uniquement. Ce problème est crucial pour les applications bioinformatiques, où il est nécessaire de considérer l'identité de séquences à un certain taux d'erreur près. C'est donc ce problème que nous avons abordé cette année.

Nous nous sommes focalisés sur le cas où les erreurs ne peuvent être que des remplacements de lettres (cas de distance de Hamming) dont le nombre est borné par K. À la différence du cas exact, une première difficulté ici consiste à donner une définition adéquate de la notion de répétition approchée. Nous avons étudié plusieurs définitions et en particulier nous en avons choisi deux qui sont d'une importance particulière, car elles représentent les cas le plus fort et le plus faible, et par conséquent, « englobent » toutes les répétitions approchées possibles. Pour chacune de ces définitions nous avons proposé un algorithme recherchant toutes les répétitions en question en temps $O(nK \log K + S)$ où S est le nombre de répétitions trouvées. En particulier, si K est considéré comme constant, nous obtenons des algorithmes linéaires de recherche. Ces résultats améliorent en particulier l'algorithme proposé par G. Landau et J. Schmidt[LS93] pour le cas de la distance de Hamming, resté jusqu'à maintenant l'algorithme de référence pour le problème. Nous avons également étudié d'autres définitions possibles de répétitions approchées et nous avons éclairci leurs rapports avec les deux définitions étudiées précédemment.

Ces résultats ont été publiés dans un rapport de recherche INRIA [10] et, dans une forme abrégée, ont été présentés à l'*European Symposium on Algorithms* à Århus, Danemark [7]. Un article décrivant ces résultats est également soumis à la revue *Theoretical Computer Science*. Notons que les algorithmes proposés ont été implantés dans le logiciel mreps décrit dans la section 5.2.

[LS93] G. Landau, J. Schmidt, « An algorithm for approximate tandem repeats », *in : Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching*, A. Apostolico, M. Crochemore, Z. Galil, U. Manber (éditeurs), *Lecture Notes in Computer Science*, 684, Springer-Verlag, Berlin, p. 120–133, Padova, Italy, 1993.

6.2 Géométrie discrète

Convexité discrète L'étude de la convexité d'une région discrète du plan se ramenant à celle de figures particulières appelées polyominos hv-convexes, nous avons développé l'année passée un algorithme incrémental et linéaire de détection de la convexité de tels polyominos[DRRRD00]. Cet algorithme parcourt les points du bord du polyomino et décide pour chaque point ajouté si le bord déjà parcouru est convexe ou non. Cette année, notre travail a porté essentiellement sur la rédaction de la preuve de la linéarité de notre algorithme. Il a donné lieu à une publication acceptée dans la revue *Discrete Applied Mathematics* [3].

Cette preuve nous a conduits à réfléchir à des aspects théoriques inattendus de l'algorithme de reconnaissance incrémentale d'un segment discret. Dans cette étude, la notion de motif géométrique reproduit périodiquement joue un rôle important. Ainsi, pour prouver la linéarité de l'algorithme, nous avons montré que, bien que nous effectuions des retours en arrière, nous ne passions le plus souvent pas plus de deux fois sur un même point, le nombre de passages multiples (au-delà de deux fois) étant lui-même majoré par le nombre de points du contour étudié. En retour, ces considérations pourront donner lieu à quelques améliorations de l'algorithme. Par exemple, en cas de retour en arrière, il est possible de limiter l'ampleur de celui-ci, à condition de retenir au cours du premier passage quelques informations complémentaires.

Nous développons actuellement une interface graphique permettant de visualiser le fonctionnement de l'algorithme, interface qui pourra être aussi utilisée pour tester des notions nouvelles de « quasi-convexité ».

Droites discrètes 3D et mesure de la longueur d'une courbe discrète En collaboration avec David Coeurjolly (Equipe de Recherche en Ingénieurie des Connaissances, Lyon II) et Olivier Teytaud (Institut des Sciences Cognitives, Bron), un algorithme de calcul de la longueur de courbes discrètes 3D a été élaboré. Il repose sur une définition arithmétique de droites discrètes 3D et sur l'algorithme linéaire de segmentation de courbes 3D donné dans [DR95]. Le principe de l'algorithme est le suivant : la courbe 3D est découpée en segments de droites de longueurs maximales et la longueur de la courbe est calculée en fonction de la longueur de la ligne polygonale obtenue. La convergence de cette technique d'estimation de longueur a été prouvée en démontrant que l'erreur entre la longueur d'une courbe, notée $\mathcal{C}$, et la longueur de cette courbe discrétisée, notée $\mathcal{C}_d$, dans une grille de taille δ est telle que :

$$|l(\mathcal{C}) - l(\mathcal{C}_d)| \leq O(\delta)$$

Ces travaux ont donné lieu à une publication dans *Lecture Notes in Computer Science* [6].

Reconnaissance de morceaux de plans discrets Le problème ici consiste à déterminer si un sous-ensemble connexe, borné, de $\mathbb{Z}^3$ est ou non un morceau de plan discret naïf. Un plan discret naïf est défini par quatre entiers μ, a, b et c et ses points (x, y, z) de $\mathbb{Z}^3$ vérifient

[DRRRD00] I. Debled-Rennesson, J.-L. Rémy, J. Rouyer-Degli, « Detection of the Discrete Convexity of Polyominoes », *in: DGCI'2000, Uppsala, Suède, Lecture Notes in Computer Science, 1953*, Springer-Verlag, p. 491–504, décembre 2000.

[DR95] I. Debled-Rennesson, *Étude et reconnaissance de droites et plans discrets*, Thèse de doctorat, Université Louis Pasteur, Strasbourg, Décembre 1995.

$\mu \leq ax + by + cz < \mu + max(|a|, |b|, |c|)$, le vecteur (a, b, c) est le vecteur normal du plan. L'algorithme de reconnaissance existant [DR96] est incrémental et reconnaît des morceaux de plans de formes rectangulaires : les sections du morceau à reconnaître, parallèles à un plan de coordonnées, sont parcourues en ajoutant les voxels un à un et en décidant par des constructions géométriques simples si l'ensemble déjà parcouru plus le point ajouté est ou non un morceau de plan discret.

L'année passée, à l'occasion de la visite de Mostefa Mesmoudi, chercheur algérien de l'Université de Mostaganem, un travail de recherche a été initié sur la mise au point d'un algorithme de reconnaissance de morceaux de plans de formes quelconques. Ce travail s'est poursuivi cette année en collaboration avec Paul Zimmermann (projet SPACES), de nouvelles techniques de construction sont en phase d'élaboration. La rédaction de ces travaux est en cours.

6.3 Aléa discret

Combinatoire analytique et analyse d'algorithmes Le travail entamé avec Cyril Banderier, Philippe Flajolet, Bruno Salvy et Michèle Soria sur l'analyse du phénomène d'Airy en combinatoire analytique s'est continué cette année. Il s'agit là de comprendre l'apparition récurrente, dans l'analyse en moyenne de différents algorithmes ou structures aléatoires discrètes, de certaines familles de lois limites non gaussiennes liées à la fonction spéciale d'Airy : alors que la loi des grands nombres conduit en général à attendre des fluctuations gaussiennes autour des valeurs moyennes, nous mettons en évidence le mécanisme mathématique sous-jacent à l'apparition de ces autres lois. Un article reprenant et étendant les résultats présentés à *ICALP'00*[BFSS00] est à paraître dans la revue *Random Structure and Algorithms* [1]. Un autre article sur l'énumération des graphes connexes est en préparation.

Les techniques de combinatoire analytique discutées ci-dessus ont été mises à profit pour l'étude d'une famille classique de la théorie des nœuds, les entrelacs alternants, dans la continuation du stage de DEA de Sébastien Kunz-Jacques sous la direction de G. Schaeffer. Les résultats de ces travaux ont été présentés à la conférence internationale *FPSAC'01* à Phoenix, Arizona [8]. (*Formal Power Series and Algebraic Combinatorics* est la conférence annuelle du domaine, et réunit environ 200 personnes chaque année.) Un article de journal est en préparation.

L'ensemble de ces travaux ont fait l'objet d'une présentation invitée aux rencontres internationales, *7th Seminar of Analysis of Algorithms*, dont le synopsis est disponible sur le Web[4]

`http://www.loria.fr/~schaeffe/Pub/Diameter`.

[4] `http://www.loria.fr/~schaeffe/Pub/Diameter`

[DR96] I. DEBLED-RENNESSON, J.-P. REVEILLÈS, « Incremental algorithm for recognizing pieces of digital planes », *in: Spie's International Symposium on Optical Science, Engineering, and Instrumentation, Technical Conference, Vision Geometry 5, Denver, USA*, 1996.

[BFSS00] C. BANDERIER, P. FLAJOLET, G. SCHAEFFER, M. SORIA, « Planar maps and Airy phenomena », *in: Automata, Languages, and Programming - ICALP'2000, Genève, Suisse*, E. W. U. Montanari, J. Rolim (éditeur), *Lecture Notes in Computer Science, 1853*, Springer, p. 388–402, juillet 2000.

Combinatoire algébrique Dans ce domaine, la collaboration suivie avec Alain Goupil (Université du Québec à Montréal) et Dominique Poulalhon (Laboratoire d'Informatique de l'X) s'est poursuivie cette année avec la participation de Sylvie Corteel (CNRS, PRISM, Versailles). À l'occasion de l'invitation de G. Schaeffer à Montréal nous avons ainsi pu mettre à jour les relations de nos précédents travaux sur les produits de classes de conjugaison dans le groupe symétrique avec les travaux de l'école russe sur les fonctions symétriques décalées (*shifted shur functions*) et avec les conjectures de Kerov. Ces conjectures font actuellement l'objet d'un intérêt tout particulier (voir les travaux récents de Philippe Biane de ENS Paris et de Richard Stanley du MIT). Deux textes sont en préparation ; l'un avec Alain Goupil et Dominique Poulalhon sur les produits de petites classes de conjugaison fait suite aux travaux que nous avions présentés à la conférence internationale *FPSAC'00* à Moscou [GPS00], l'autre avec Alain Goupil et Sylvie Corteel traite des liens entre une base de fonctions symétriques que nous avons introduite, les *content power sums*, et les conjectures de Kerov.

Enfin notons que l'article *Factorisations of a n-cycle into m permutations* de Dominique Poulalhon et G. Schaeffer a été accepté pour publication dans la revue *Discrete Mathematics* [5].

Topologie des surfaces aléatoires Les cartes planaires et triangulations aléatoires sont un modèle combinatoire classique sur lequel nous travaillons depuis plusieurs années. Nous avons d'une part continué nos progrès dans la compréhension de ces structures et d'autre part, commencé l'exploration plus systématique des liens entre notre point de vue et les travaux conduits sur ces mêmes modèles en physique quantique.

Avec Philippe Chassaing (professeur à l'IECN), nous avons entrepris de démontrer une conjecture sur le diamètre des cartes aléatoires, formulée il y a déjà plusieurs années par G. Schaeffer. Nos résultats s'appuient sur un codage de ces objets par des arbres étiquetés et sur un passage à la limite continue inspiré des travaux du probabiliste D. Aldous. Nous travaillons plus généralement à la rédaction d'un article sur les propriétés des distances dans les cartes aléatoires. Remarquons encore que le codage précédemment mentionné a été étendu avec Michel Marcus en genre supérieur, ce qui a donné lieu à la rédaction d'un rapport de recherche ([11]).

Un aspect différent de la topologie de ces surfaces, à savoir la présence de petits séparateurs, fait l'objet d'une collaboration avec Jason Gao (Professeur à Carlton University). Cette collaboration, entamée l'an dernier, s'est poursuivie par la visite de G. Schaeffer à Ottawa. Nous avons à cette occasion obtenu quelques premiers résultats tangibles sur la distribution du nombre de séparateurs minimaux, qui devraient faire l'objet d'un article.

Le lien de ces travaux avec la physique quantique peut sembler surprenant, mais l'explication suivante, quoique très simplifiée, permet d'en appréhender l'origine. De même qu'en informatique on est naturellement amené par des contraintes matérielles à considérer des géométries discrètes (cf. la section 6.2), la recherche de modèles mathématiques adaptés à la physique débouche sur des modèles discrets. Ainsi, pour la physique statistique classique, la

[GPS00] A. Goupil, D. Poulalhon, G. Schaeffer, « Central Characters and Conjugacy Classes in the Symmetric Group », *in : Formal Power Series and Algebraic Combinatorics, Moscou, Russia*, D. Krob, A. Mikhalev, A. Mikhalev (éditeurs), Springer, p. 238–249, juillet 2000.

discrétisation de l'espace euclidien usuel à deux dimensions conduit naturellement à l'étude de modèles sur une grille régulière, tandis que, pour la physique quantique qui remplace l'univers fixe par une distribution de probabilité sur tous les univers possibles, la discrétisation conduit à une distribution de probabilité sur tous les univers discrets possibles, qui se trouve coïncider avec le modèle de surfaces aléatoires étudié en combinatoire. La conférence *Discrete Random Geometry and Quantum Gravity*[5] illustre bien ce courant de la physique quantique.

Alors que les aspects géométriques des surfaces aléatoires ont été largement étudiés aussi bien en combinatoire qu'en physique par des méthodes et avec des résultats complémentaires, les modèles sur cartes ont plus spécifiquement été considérés en physique. Avec Mireille Bousquet-Mélou, nous travaillons à comprendre comment les outils de l'approche combinatoire (décompositions des structures, séries génératrices et équations fonctionnelles) s'étendent et s'adaptent à ces problèmes. Ce travail est aussi motivé par le fait que les équations que nous avons mises à jour sont des équations aux variables catalytiques. Ces équations sont des équations algébriques en des séries algébriques inconnues faisant intervenir des évaluations de ces séries en des points particuliers des variables dites catalytiques. Leur étude, et en particulier l'étude de leurs solutions proprement algébriques, est au centre de nos travaux de ces dernières années. Remarquons sur ce dernier point que notre article, dans lequel une classe de telles équations est traitée, vient d'être accepté pour publication dans la revue *Probability Theory and Related Fields* [2].

L'ensemble de ces travaux sur la topologie des surfaces aléatoires a fait l'objet d'exposés invités au *MSRI hot topic workshop : Critical Percolation and Conformally Invariant Processes*, au séminaire du groupe *Theory* de Microsoft Research et au *First Montreal-Ottawa Analysis of Algorithms Day* à Carlton University.

Algorithmes de génération aléatoire Les algorithmes de génération aléatoire restent au centre de nos intérêts. Nous travaillons d'une part au développement de nouveaux algorithmes dédiés à des familles de graphes particulièrement intéressantes. Ainsi avec Dominique Poulalhon (LIX), nous avons pu étendre le paradigme de génération aléatoire par conjugaison d'arbres développé par G. Schaeffer à la classe importante des triangulations avec bords. Le nouvel algorithme, de complexité linéaire, unifie ainsi le cas classique des triangulations de polygones et celui des triangulations de la sphère. Un article décrivant ces résultats est en cours de rédaction.

D'autre part, nous nous intéressons à la compréhension des mécanismes généraux qui permettent le développement d'un algorithme efficace de la génération aléatoire. Dans cette direction, nous travaillons plus particulièrement à l'utilisation d'algorithmes probabilistes combinatoires : ces algorithmes combinent les avantages de l'approche combinatoire (essentiellement la garantie de respecter parfaitement la distribution probabiliste visée) avec la simplicité d'implantation des méthodes probabilistes. L'originalité de notre approche est de ne pas relâcher la contrainte d'uniformité (contrairement aux chaînes de Markov dont la convergence vers la distribution uniforme ne peut être controlée que par des méthodes complexes de couplage par le passé), mais plutôt de travailler sur la taille des objets considérés, en autorisant de légères fluctuations de ce paramètre. L'article écrit sur ce sujet en collaboration avec Philippe Duchon,

[5] `http://www1.phys.uu.nl/Symposion/EUWORKLOLL/EUWorkshop.htm`

Philippe Flajolet et Guy Louchard est soumis à la conférence *STOC'2002* [9].

Enfin nous avons entrepris de mettre à profit nos algorithmes de génération aléatoire pour étudier expérimentalement les propriétés des surfaces aléatoires. Ces travaux sont menés en collaboration avec David B. Wilson (Microsoft Research, Seattle) pour l'étude des modèles de physique statistique sur surfaces aléatoires gelées, et avec Igor Rivin (Princeton) sur les propriétés spectrales des triangulations aléatoires.

Combinatoire des suites arithmétiques L'année 2001 a par ailleurs correspondu à l'achèvement d'un article sur les suites arithmétiques autodécrites, à paraître dans la revue *Advances in Applied Mathematics* [4].

7 Actions régionales, nationales et internationales

7.1 Actions régionales

Au niveau du Contrat de Plan Etat-Région 2000-2006, nous sommes impliqués dans le Pôle de Recherche Scientifique et Technologique (PRST) *Intelligence Logicielle*.

L'équipe ADAGE participe au

- Génopôle Strasbourg Alsace-Lorraine
- Thème « Bioinformatique et Applications à la Génomique » du PRST Intelligence Logicielle
- Thème « Qualité et sûreté des logiciels et systèmes informatiques » du PRST Intelligence Logicielle

En outre, nous avons des collaborations soutenues avec des mathématiciens de l'Institut Elie Cartan de Nancy : Ph. Chassaing sur le thème de l'aléa discret (voir la section 6.3) et P. Vallois et M.-P. Etienne sur le thème de la bioinformatique.

7.2 Actions nationales

L'équipe participe activement au groupe Aléa, composante du GDR CNRS ALP. Nous avons aussi participé à la réponse à l'appel d'offre Math-STIC du CNRS au sein du groupe coordonné par Brigitte Chauvin et Brigitte Vallée.

Nous participons au projet GÉNOGRID, mis en place cette année dans le cadre de l'Action Concertée Incitative *Globalisation des ressources informatiques et des données* (ACI GRID). Le projet est coordonné à l'IRISA par Dominique Lavenier, les autres laboratoires participants sont LAMIH, ABISS, LIH, LIFL.

ADAGE participe à l'action CNRS *Algorithmes pour la bioinformatique (ALBIO)* qui vient de se créer. G. Kucherov est co-responsable du thème *Séquences répétées* dans le cadre de cette action.

Nous avons participé au projet « Approches multicritères pour la modélisation et l'analyse in silico des génomes », qui s'est constitué en 2001 en réponse à l'appel d'offres commun CNRS - INRA - INRIA - INSERM (9 laboratoires français d'informatique et de biologie impliqués, projet accepté pour un an).

7.3 Actions internationales

Des membres d'ADAGE participent au projet INTAS *Methods, algorithms and software for functional and structural annotation of complete genomes* (projet avec la Russie, incluant des laboratoires d'Allemagne, de France et d'Autriche).

G. Schaeffer collabore avec David B. Wilson de Microsoft Research à Seattle, Jason Gao à l'université Carlton d'Ottawa et Alain Goupil de l'université du Québec à Montréal. Ces liens se sont concrétisés par des séjours dans ces universités et la préparation d'articles en commun.

Bien que notre projet dans le cadre de l'Institut Liapunov franco-russe d'Informatique et de Mathématiques appliquées soit arrivé à son terme l'année dernière, nous gardons des contacts étroits avec les chercheurs moscovites, ce qui s'est concrétisé cette année par le séjour de R. Kolpakov dans l'équipe et notre travail commun.

7.4 Visites, et invitations de chercheurs

Roman Kolpakov, chercheur de l'Université de Moscou et collaborateur de l'Institut Liapunov franco-russe d'Informatique et Mathématiques Appliquées, a travaillé au sein d'ADAGE huit mois en 2001, dans le cadre de son contrat INRIA d'accueil de spécialistes.

Mohammed Tajine, professeur d'informatique à Strasbourg, a fait une visite de deux jours chez ADAGE au mois d'avril.

Jean-Pierre Reveillès, professeur à Clermont-Ferrand et directeur du LLAIC, est venu en visite trois jours dans l'équipe en juin et a fait un exposé au séminaire d'informatique fondamentale du LORIA, intitulé *Courbure discrète*.

Alain Daurat est venu deux jours en mars et a fait un exposé au séminaire d'ADAGE intitulé *Reconstruction des parties convexes du plan discret à partir de leurs projections*.

Ania Gambin, professeur de l'Université de Varsovie (Pologne), a été invitée pour un mois pour travailler sur le thème de l'alignement de protéines. Elle a fait un exposé au séminaire de bioinformatique du Loria.

Mireille Bousquet-Mélou, CR CNRS du LaBRI, a séjourné au LORIA une dizaine de jours fin septembre.

8 Diffusion de résultats

8.1 Animation de la Communauté scientifique

Tous les membres d'ADAGE se sont réunis le 2 mars pour une « journée au vert » lors de laquelle ils ont fait une présentation de leurs travaux, suivie d'une discussion sur les perspectives. D'autres membres du Loria y ont été invités.

I. Debled-Rennesson est membre de la commission de spécialistes de l'IUFM de Lorraine.

L'année passée, G. Kucherov a fait partie des comités de programme des conférences *Journées Ouvertes : Biologie, Informatique et Mathématiques (JOBIM'2001)* et *Perspectives of System Informatics (PSI'2001)*. Il fait actuellement partie des comités de programme de *JOBIM'2002* et des *Journées d'Arithmétiques Faibles (JAF'2002)*.

J. Rouyer est membre du bureau de la commission de spécialistes 27ème section de l'Université Henri Poincaré.

G. Schaeffer a organisé avec Philippe Chassaing de l'IECN les rencontres ALÉA 2001 au CIRM à Marseille. Il est également responsable du séminaire d'informatique fondamentale du LORIA.

8.2 Enseignement universitaire

G. Kucherov a soutenu son mémoire d'habilitation à diriger les recherches [Kuc00].

G. Kucherov, en commun avec D. Kratsch (Université de Metz), enseigne le module *Algorithmique des structures discrètes* du DEA d'Informatique à Nancy (filière *Algorithmique Numérique et Symbolique*). Il dispense également des cours en DESS *Ressources Génomiques et Traitements Informatiques* à l'Université Henri Poincaré de Nancy.

G. Kucherov, I. Debled-Rennesson et G. Schaeffer ont encadré le stage d'Alexeï Stanger (stage de deuxième année du Magistère d'informatique de l'Universite de Paris 6) pendant mai-septembre 2001.

G. Kucherov a encadré le stage de 6 semaines d'Emmanuelle Becker (stage de première année du Magistère d'informatique de l'ENS de Lyon), le stage de Ralph Rabbat (*Massachusetts Institut of Technology*), et le projet tutoré d'Émilie Testa (DESS de Ressources Génomiques et Traitements Informatiques, Université Henri Poincaré Nancy 1)

J. Rouyer a encadré un projet de découverte de la recherche de l'Ecole Supérieure d'Informatique et Applications de Lorraine d'octobre à décembre 2001. Elle est par ailleurs responsable de la spécialisation Ingénierie du Logiciel de l'ESIAL.

G. Schaeffer a fait deux enseignements à l'École Normale Supérieure de Cachan : il a donné début janvier une journée de conférences (6 heures) dans le cadre des conférences du département d'informatique ; il a également été invité à intervenir dans le cursus d'informatique de l'École Normale Supérieure de Cachan (une journée de conférence et un cours (32h) sur le thème *Combinatoire et théorie des graphes*).

8.3 Participation à des colloques, séminaires, invitations

8.3.1 Colloques, tutoriels, conférences et séminaires invités

I. Debled-Rennesson a participé à l'école d'hiver "Digital and Image Geometry" qui s'est déroulée pendant une semaine au château de Dagstuhl en Allemagne.

I. Debled-Rennesson a fait un exposé, en collaboration avec B. Aigle (Laboratoire de Génétique et de Microbiologie, UHP), en novembre, à la journée *Bioinformatique et Applications à la Génomique*, organisée au LORIA dans le cadre du PRST *Intelligence Logicielle*.

Au mois de février, G. Kucherov s'est rendu aux États-Unis pour visiter le groupe de bioinformatique de la société *GlaxoSmithKline* près de Philadelphie, où il a fait un exposé. Pendant ce voyage il a également visité la société *Compugene Inc.* dans l'état de *New Jersey*.

G. Kucherov a fait un séminaire invité au LIFAR (Laboratoire d'Informatique Fondamentale et Appliquée de Rouen) au mois d'avril. Il a également fait un exposé à la conférence *Logic*

[Kuc00] G. Kucherov, *Motifs dans les mots et les arbres*, Habilitation à diriger des recherches, Université Nancy 1 Henri Poincaré, Decembre 2000.

and Complexity in Computer Science (LCCS'2001) à Créteil en septembre 2001, et un exposé au groupe de travail de bioinformatique AMASIG réuni au LRI à Orsay en novembre.

G. Kucherov et R. Kolpakov ont présenté leur travail à l'*European Symposium on Algorithms (ESA'2001)*, qui a eu lieu au Danemark au mois d'août dans le cadre de la fédération de conférences ALGO'2001. Ils ont également participé, dans le cadre d'ALGO'2001, au *Workshop on Algorithms in Bioinformatics (WABI)*.

G. Kucherov a participé aux manifestations scientifiques suivantes :

- journée de la génomique organisée au mois de mars par l'Institut Elie Cartan de Nancy,
- conférence *RECOMB'01* à Montréal au mois d'avril,
- conférence *JOBIM'01* à Toulouse au mois de mai,
- conférence *Perspectives of System Informatics (PSI'01)* à Novossibirsk (Russie) au mois de juillet

J.-L. Rémy a participé à la rencontre ALEA au CIRM à Marseille-Luminy.

J. Rouyer et R. Kolpakov ont participé à la conférence JOBIM'01 à Toulouse au mois de mai.

G. Schaeffer a fait les exposés invités suivants :

- exposé à la conférence *MSRI Hot Topics workshop : Critical Percolation and Conformally Invariant Processes* en mai 2001 à Berkeley,
- exposé aux rencontres *First Montreal-Ottawa Analysis of Algorithms Day* en juin 2001 à Ottawa,
- exposé long (1h15) au colloque international *7th Seminar on Analysis of Algorithm* en juillet 2001 à Tatihou,
- exposé au séminaire du laboratoire d'informatique de l'Université Marne-la-Vallée,
- exposé au séminaire de l'équipe *Theory* des laboratoires *Microsoft Research* à Seattle.

Il a également fait un exposé aux ateliers *GASCOM'2001* à Sienne en novembre 2001.

Les résultats de G. Schaeffer et Sébastien Kunz-Jacques ont donné lieu à un exposé à la conférence *13th International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC'01)* à Phoenix, Arizona en juin 2001.

G. Schaeffer a participé au colloque international *RECOMB'01* à Montréal et aux rencontres *Discrete Random Geometries and Quantum Gravity* à Utrecht.

8.3.2 Séjours de chercheurs

G. Schaeffer a été invité une semaine en mai dans les laboratoires de *Microsoft Research* à Seattle, au sein de l'équipe *Theory*, où il collabore avec David B. Wilson, et une semaine en juin à l'Université Carlton d'Ottawa, à l'initiative du professeur Jason Gao. Il a séjourné deux mois au printemps au sein du LaCIM à l'université du Québec à Montréal à l'invitation d'Alain Goupil.

En France, G. Schaeffer s'est rendu à Bordeaux en juillet pendant deux semaines pour y travailler au LaBRI avec Mireille Bousquet-Mélou, ainsi qu'à plusieurs reprises à l'IGBMC à Strasbourg, dans le cadre de notre collaboration au sein du Génopole, pour y travailler particulièrement avec Frédéric Plewniak et Olivier Poch. G. Kucherov s'est également rendu à Strasbourg dans le même cadre au mois d'octobre.

8.4 Jurys de thèses et jurys divers

I. Debled-Rennesson a été membre du jury de magister de Mlle Boukhatem, étudiante au département de mathématiques de la faculté des sciences de l'Université de Mostaganem (Algérie).

G. Kucherov a été rapporteur de la thèse de doctorat de Cyril Allauzen (Université Marne-la-Vallée).

G. Schaeffer fait partie du jury du prix de thèse annuel de l'association SPECIF, après avoir reçu ce prix pour l'année 1999.

9 Bibliographie

Articles et chapitres de livre

[1] C. Banderier, P. Flajolet, G. Schaeffer, M. Soria, « Random Maps, Coalescing Saddles, Singularity Analysis, and Airy Phenomena », *Random Structures and Algorithms*, 2001, p. 1–47, à paraître.

[2] M. Bousquet-Mélou, G. Schaeffer, « Walks on the slitplane », *Probability Theory and Related Fields*, 2001, p. 1–30, à paraître.

[3] I. Debled-Rennesson, J.-L. Rémy, J. Rouyer-Degli, « Detection of the Discrete Convexity of Polyominoes », *Discrete Applied Mathematics*, 2001, accepté pour publication.

[4] Y. Pétermann, J.-L. Rémy, I. Vardi, « Discrete Derivatives of Sequences », *Advances in Applied Mathematics 27*, 2001, p. 562–584.

[5] D. Poulalhon, G. Schaeffer, « Factorizations of large cycles in the symmetric group », *Discrete Mathematics*, 2001, p. 1–26, sous presse.

Communications à des congrès, colloques, etc.

[6] D. Coeurjolly, I. Debled-Rennesson, O. Teytaud, « Segmentation and Length Estimation of 3D Discrete Curves », *in : Digital and Image Geometry*, A. I. G. Bertrand, R. Klette (éditeurs), *Lecture Notes in Computer Science, 2243*, Springer, p. 295–313, 2001.

[7] R. Kolpakov, G. Kucherov, « Finding Approximate Repetitions under Hamming Distance », *in : 9-th European Symposium on Algorithms (ESA 2001), Aarhus, Denmark*, F. auf der Heide (éditeur), *Lecture Notes in Computer Science, 2161*, p. 170 – 181, août 2001.

[8] S. Kunz-Jacques, G. Schaeffer, « The asymptotic number of prime alternating links », *in : Formal Power Series and Algebraic Combinatorics, Phenix, Arizona*, 2001.

Rapports de recherche et publications internes

[9] P. Duchon, P. Flajolet, G. Louchard, G. Schaeffer, « Random Sampling from Boltzmann principles », *rapport de recherche*, LORIA, 2001, 12pp, soumis.

[10] R. Kolpakov, G. Kucherov, « Finding Approximate Repetitions under Hamming Distance », *Rapport de recherche*, INRIA, avril 2001, `http://www.inria.fr/rrrt/rr-4163.html`.

[11] M. Marcus, G. Schaeffer, « Une bijection simple pour les cartes orientables », *rapport de recherche*, LORIA, 2001, 10pp.