

UNIVERSITÉ PAUL SABATIER TOULOUSE III

U.F.R. MATHÉMATIQUE INFORMATIQUE GESTION

THESE

POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ TOULOUSE III

Discipline : **Mathématiques appliquées**

Présentée et soutenue par

Peggy CÉNAC

le 13 Juin 2006

**Étude statistique de séquences biologiques
et convergence de martingales**

Directeurs de thèse : Bernard BERCU et Guy FAYOLLE

Jury

M. Bernard BERCU,	Directeur,
M. Guy FAYOLLE,	Directeur,
M. Philippe FLAJOLET,	Président,
M. Fabrice GAMBOA,	Examineur,
M. Didier PIAU,	Rapporteur,
M. Alexander TSYBAKOV,	Rapporteur.

Thèse réalisée au sein du projet PRÉVAL, INRIA Rocquencourt

Remerciements

J'emprunte à Antoine de Saint Exupéry ces mots éclairés : « La vérité de demain se nourrit de l'erreur d'hier, et [...] les contradictions à surmonter sont le terreau même de notre croissance. » Je souhaiterais sincèrement remercier toutes les personnes sans lesquelles cette thèse n'aurait jamais pu voir le jour.

Je tiens tout d'abord à exprimer ma reconnaissance et mon admiration à deux personnes que je ne peux dissocier, tant pour leur disponibilité, leur gentillesse, que pour leur dynamisme, sources d'encouragements et de bonne humeur. Guy, ta porte était toujours ouverte pour recevoir mes questions et dissiper mes doutes ; Bernard, tu as su me donner le goût de la recherche en DEA, et tu as réussi à faire en sorte que Toulouse reste toujours proche de Rocquencourt. Pour ces raisons et bien d'autres encore, je vous remercie.

Didier Piau et Alexander Tsybakov ont accepté la lourde tâche de rapporter ma thèse, sans se laisser rebuter par la diversité des sujets. Fabrice Gamboa (à qui je suis reconnaissante de m'avoir présenté Guy) et Philippe Flajolet me font l'honneur d'accepter d'être membres de mon jury. Je leur adresse toute ma gratitude.

Je remercie les autres habitants du bâtiment 20 (Jean-Marc Lasgouttes, Arnaud de La Fortelle, Cyril Furtlehner, Philippe Le Chenadec) pour m'avoir chaleureusement accueillie. Je garde le souvenir de discussions toujours enrichissantes. Merci aussi à Nicolas Gibelin pour les longueurs matinales partagées dans la piscine du Chesnay.

Je remercie également le personnel des moyens informatiques de l'INRIA Rocquencourt, notamment l'équipe AFS d'Edmonde Duteurtre ainsi que l'équipe du cluster, ressource précieuse pour nos expérimentations. Il est parfois facile d'oublier que si tout fonctionne, c'est justement parce qu'ils sont là.

Brigitte, Nicolas et Stéphane, travailler avec vous, toujours dans la bonne humeur, a fait de notre collaboration, un moment privilégié.

Je tiens également à remercier la conviviale équipe de Mathématiques de la faculté Jean Monnet (Farida Malek, Luc Joseph, Patrick Beau, Nicolas Bousquet), qui a su guider mes premiers pas dans l'enseignement. Je suis particulièrement reconnaissante à ma tutrice Odile Brandière de s'être montrée toujours disponible chaque fois que j'avais besoin d'elle.

Mon éveil aux mathématiques et à la rigueur scientifique est dû en grande partie à Claudine Pelisse. Je la remercie. Mes remerciements vont aussi à d'autres enseignants qui ont contribué, pour beaucoup, à mon goût des mathématiques : Thomas Lafforgue pour la confiance qu'il m'a témoignée et son soutien constant, Wendelin Werner pour

la découverte captivante des probabilités, Dominique Hulin et Jean-Christophe Léger qui ont su, par leurs qualités tant humaines que pédagogiques, préserver ma motivation dans un contexte difficile de préparation à l'agrégation.

Jean-Marc, ton oreille attentive, tes nombreuses relectures et tes remarques constructives, ta présence, ont été pour moi une aide précieuse. J'adresse donc toute ma gratitude à mon voisin de bureau.

Merci à ma famille et mes amis d'avoir accepté mon rythme de travail, mes absences et mes silences. Parce que votre amitié m'a aidée plus que vous ne le pensez, merci à Magalie, Claire, Olivier, Chrisline, Marie, Manue, David, Marie et Nico. Un merci tout particulier à Fred. Je suis également reconnaissante à tous les artistes du Lapin, et à Vincent, pour toutes les bulles d'oxygène qu'ils m'ont données.

Je remercie mes parents d'avoir choisi de m'élever dans un monde empreint de curiosité scientifique, de m'avoir toujours encouragée et soutenue dans toutes mes études. Parce que je leur dois tant, je leur dédie cette thèse.

Enfin, je suis reconnaissante à Maxence, dont l'écoute parfaite et la main toujours tendue, m'ont soutenue aussi bien par ses doubles-journées de travail pour le développement de nos programmes informatiques que pour la relecture consciencieuse du manuscrit. Merci d'être le compagnon inespéré que le destin m'a donné.

Les sentiers tortueux de l'hominisation

L'homme n'est pas seulement un être de culture, il est aussi le produit d'une évolution naturelle, qui, au fil des millions d'années, avec une multitude d'autres facteurs extérieurs, a fait qu'il est devenu l'homme.

Ce processus d'évolution ne repose ni sur un enchaînement de hasards ni sur une prédestination. La vieille querelle sur les origines de l'homme s'est égarée. Mieux nous analysons le déroulement de notre phylogenèse, mieux nous comprenons nos caractères spécifiques et nos problèmes, et plus nous devrions être conscients des liens étroits qui nous unissent avec la nature dont nous sommes issus et dont nous continuons à faire partie intégrante. Le fait que nous soyons aujourd'hui en mesure de mieux comprendre nos origines nous contraint à nous montrer plus responsables non seulement à l'égard de nos semblables, mais encore envers la nature. Peut-être cette crise de l'environnement que nous traversons aujourd'hui résulte-t-elle de ce que trop d'entre nous ont méprisé leurs origines et regardé la nature de haut. Renier ce dont nous sommes issus ne nous rend pas meilleurs - le reconnaître ne nous rend pas mauvais. Peut-être pouvons-nous simplement arriver à une meilleure compréhension de nous-mêmes en tenant compte des conditions extérieures dans lesquelles nous sommes apparus.

Josef Reichholf
L'émergence de l'homme

A mes parents

Avant-propos

« *La science ne sert guère qu'à nous donner
une idée de l'étendue de notre ignorance.* »

Félicité de Lamennais

Dans sa course à la connaissance et à la compréhension du phénomène de la vie, l'Homme a franchi plusieurs étapes décisives au siècle dernier, avec notamment la découverte de la structure de *l'acide désoxyribonucléique* (ADN), par Crick et Watson en 1953. Cette molécule contient l'information à partir de laquelle sont construits les organismes vivants, constitués pour l'essentiel de *protéines* déterminées.

L'ADN constitue le ou les chromosomes que l'on trouve dans les cellules. C'est une molécule composée de deux brins, chacun orienté par un sens de lecture opposé, formant la fameuse double hélice. Ces brins sont composés de quatre bases azotées, les *nucléotides* : adénine (A), cytosine (C), guanine (G), thymine (T). Une relation de complémentarité existe entre l'adénine et la thymine, d'une part, et entre la cytosine et la guanine d'autre part. Chaque nucléotide d'un brin est associé à son complémentaire sur l'autre brin.

L'enchaînement des lettres (nucléotides) le long des brins permet la synthèse d'acides aminés. Il en existe 20, chaque triplet de nucléotides correspondant à un acide aminé. Comme le nombre de triplets dans un alphabet de 4 lettres est très largement supérieur à 20, plusieurs triplets codent le même acide aminé : on dit que le code génétique est *redondant*. Les protéines sont des hétéropolymères d'acides aminés, c'est-à-dire des agrégats d'acides aminés différents. Les séquences de nucléotides sur les brins d'ADN permettent donc la synthèse de différentes protéines. Cependant, seules certaines parties des brins sont *codantes*. En effet, les molécules, qui agissent comme catalyseurs des réactions chimiques à l'origine de la synthèse des acides aminés et donc des protéines, nécessitent des conditions particulières pour agir. Citons par exemple la forme du brin, les suites de nucléotides pour signaler le début et la fin d'un code de protéine, la température. Les zones qui *s'expriment* ainsi sont appelées *exons*, un gène correspondant à un ou plusieurs exons. Les *gènes* sont des parties de séquences d'ADN qui *codent* pour la production de protéines spécifiques.

Le développement des puces à ADN, dans les années 1990, a permis à l'Homme de *séquencer* différentes espèces dont la sienne, c'est-à-dire d'obtenir la succession des bases dans leurs séquences d'ADN. Deux individus d'une même espèce n'ont certes pas des molécules d'ADN identiques, mais ces différences, bien que visibles, sont négligeables par rapport au fonctionnement commun à toute l'espèce considérée.

Avec le séquençage venait l'espoir de comprendre les fonctionnements et l'information

contenue dans l'ADN, afin de saisir les mécanismes à l'origine de la vie, de l'évolution, du vieillissement, ou encore soigner certaines pathologies génétiques.

Cependant, la masse d'information accumulée, suite aux différents séquençages d'espèces, a soulevé de nombreux problèmes de stockage : les principales banques de séquences contiennent depuis août 2005 plus de 100 milliards de paires de bases. Il est donc devenu urgent et nécessaire de développer des outils permettant de retrouver rapidement l'information *pertinente* dans cet amas de données. La proportion des exons dans la totalité des séquences d'une espèce est souvent très faible : par exemple pour l'Homme, on estime à 1,5% la quantité d'ADN codant. Trouver des moyens de repérer les gènes fait partie des nombreuses recherches effectuées à l'heure actuelle en bioinformatique, tout comme la compréhension de leurs interactions. En effet, il est rare qu'un seul gène soit à l'origine d'un mécanisme donné ; si c'était le cas, une simple mutation sur ce gène pourrait mettre en péril tout le mécanisme en question.

La bioinformatique est une discipline située entre l'informatique et la biologie. Elle a pour but de fournir des outils adaptés aux analyses biologiques. Ces dernières années, cette discipline a donc connu un essor considérable. Les méthodes statistiques sont également à l'origine de méthodes efficaces pour trouver des sites intéressants sur l'ADN, par exemple en signalant des mots exceptionnellement rares ou fréquents, qui peuvent signaler une fonction spéciale de la partie d'ADN correspondante.

C'est dans ce contexte que peut-être utilisée la Chaos Game Representation (CGR). Il s'agit d'une méthode de stockage et surtout d'une méthode de représentation graphique de séquences, appliquée pour la première fois aux séquences d'ADN par Jeffrey [47]. Le but de la première partie de cette thèse est d'apporter des éléments de réponse à la question suivante : comment utiliser la CGR et l'information qu'elle contient ?

La visualisation de l'ADN sous cette forme graphique permet de comparer des motifs, extraits de séquences, localement comme globalement. La CGR est un système dynamique qui, à une séquence de lettres dans un alphabet fini, fait correspondre une trajectoire dans un espace continu, voire une mesure empirique sur un ensemble. Comment utiliser une telle mesure pour comparer deux séquences biologiques de façon pertinente ?

L'une des propriétés fondamentales de la CGR est que chaque point contient toute l'histoire de la séquence parcourue. Quel est le gain d'information de la CGR par rapport aux méthodes classiques basées sur les comptages de mots ?

Les chapitres 1 et 2 sont le prolongement des travaux [17] réalisés en collaboration avec G. Fayolle et J.-M. Lasgouttes. Dans le Chapitre 1, on commence par définir la construction de la CGR et énoncer ses principales propriétés. On établit ensuite un rapide état des lieux de ses applications.

Le Chapitre 2 est consacré à l'étude de nouvelles applications de cette représentation. Si la séquence à analyser est stationnaire, alors la suite des points issus de la CGR est une chaîne de Markov d'ordre un, quel que soit l'ordre de dépendance de la séquence à analyser. À partir de propriétés sur la mesure invariante des points de la CGR, il est possible de déterminer la structure de la séquence elle-même. Cette caractérisation

permet de construire une famille de tests pour déterminer l'ordre d'une chaîne de Markov homogène représentant l'évolution de la séquence. Les zones de rejet sont basées sur des partitions de l'espace d'arrivée de la CGR, et les statistiques utilisées ne font pas obligatoirement intervenir le comptage de mots de la séquence [15]. Ainsi, pour une séquence donnée, on détermine avec ces tests un ordre de dépendance et on peut ensuite lui associer un modèle. Le modèle choisi pourra par exemple servir de référence pour juger de la pertinence des écarts à une situation moyenne, pour trouver des mots de fréquences « exceptionnelles » dans une séquence biologique donnée. En effet, un modèle markovien d'ordre d permet de prendre en compte les fréquences de mots de longueur d et d'évaluer la significativité des occurrences des mots de longueur $d + 1$.

La deuxième application présentée au Chapitre 2 est une généralisation, au moyen de la CGR, de la notion de profils d'abondance relative de dinucléotides. Ces profils sont utilisés par Karlin et Burge [50], Campbell et al. [13], Jernigan et Baran [48], pour définir une signature génomique. L'action de l'environnement et des systèmes de maintenance de l'ADN (réplication, réparation, recombinaison) influent sur la structure des génomes. L'ensemble des fréquences des mots de différentes longueurs peut être considéré comme une caractéristique du génome. Avec la CGR, on définit un nouveau profil. On souligne les performances de la CGR pour classifier des espèces et construire des arbres taxonomiques par rapport aux techniques précédentes à base de comptage de mots.

Puisque la CGR permet de visualiser les fréquences de tous les suffixes d'une séquence d'ADN, il est alors naturel de lui associer un arbre quaternaire, dont la construction permet de grouper ces répétitions de suffixes.

Le travail présenté dans le Chapitre 3 est issu d'une collaboration avec B. Chauvin, N. Pouyane et S. Ginouillac [16]. On y définit l'*arbre-CGR*. Ce dernier pousse en insérant successivement les préfixes d'une séquence. Plus précisément, à partir d'une séquence de lettres, on construit un arbre digital de recherche (*Digital Search Tree* ou *DST*) par insertion successive des préfixes retournés de la séquence. L'arbre avec étiquettes numérotées est équivalent à la CGR, tandis que la donnée d'un arbre sans étiquettes est équivalente à une liste non ordonnée de mots présents dans la séquence.

Bien que la construction de ces arbres soit inspirée de la CGR, ils peuvent néanmoins être obtenus directement à partir de la séquence. Cependant, cette construction est motivée par la volonté de mesurer de nouvelles quantités statistiques, cachées dans la séquence mais visibles sur l'arbre, afin de dégager de nouvelles caractéristiques pour une loi de génération donnée.

Plusieurs résultats sont connus sur la hauteur, la profondeur d'insertion et le profil, pour des DST construits sur des suites de séquences indépendantes, générées par une source émettant des tirages indépendants et de même loi (*i.i.d.*). En particulier, pour le modèle de Bernoulli, les arbres sont binaires et les séquences insérées sont indépendantes les unes des autres ; de plus, chaque séquence est une suite de variables aléatoires *i.i.d.* et les deux lettres sont équiprobables. Dans ce contexte, la littérature est abondante.

Cependant, si les DST sont construits sur des séquences, émises par des sources indé-

pendantes et identiquement distribuées mais biaisées ou markoviennes, seul Pittel [74] obtient des résultats de convergence sur la profondeur d'insertion et sur la hauteur.

Nous montrons dans le Chapitre 3 que la différence de comportement asymptotique entre les arbres-CGR et les DST classiques construits à partir de séquences markoviennes n'est pas visible au premier ordre.

Le Chapitre 4 est une extension de l'article [14]. On y propose de nouvelles lois fortes des grands nombres pour les martingales vectorielles avec diverses applications statistiques. Le contexte est le suivant.

Soit (ε_n) une suite de variables indépendantes centrées et de même loi, de variance σ^2 . En notant $Z_n \stackrel{\text{def}}{=} \varepsilon_1 + \dots + \varepsilon_n$, le théorème dit *de la limite centrale presque sûr*, établit que, pour toute fonction h continue bornée, on a avec probabilité 1,

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n \frac{1}{k} h\left(\frac{Z_k}{\sqrt{k}}\right) = \int_{\mathbb{R}} h(x) dG(x),$$

où G est la mesure gaussienne $\mathcal{N}(0, \sigma^2)$. On peut trouver une version martingale de ce théorème dans Chaâbane et Maâouia [20]. Les questions suivantes sont alors naturelles : le théorème reste-t-il vrai pour des fonctions h non bornées et qu'en est-il du cadre vectoriel ? On propose ici de nouvelles propriétés asymptotiques presque sûres pour les moments de transformées de martingales vectorielles. On établit ainsi des résultats de convergence sur les erreurs d'estimation et de prédiction cumulées, associées à certains modèles de régression tels les modèles autorégressifs linéaires et les processus de branchement avec immigration.

La thèse fait usage de divers résultats classiques de convergence sur les martingales, mais aussi de méthodes analytiques, principalement d'analyse complexe, notamment pour l'étude de certaines transformées.

Enfin, dans le but de mettre en pratique les différents résultats théoriques, une suite de programmes a été développée en Objective-Caml, en collaboration avec M. Guesdon. Elle inclut une bibliothèque de modules et une interface graphique. Ces programmes sont mis à disposition sur le site <http://mycgr.inria.fr/>. La description détaillée des programmes apparaît au cours des différents chapitres de ce manuscrit.

Notations

– Ensembles

$\mathcal{B}(S)$: ensemble des boréliens de l'ensemble S .

$|S|$: cardinal de l'ensemble S .

$|w|$: longueur du mot w , i.e. $|w_1w_2 \dots w_n w| = n$.

$\dim_H(S)$: dimension de Hausdorff de S .

– Fonctions

$\Re(f)$: partie réelle de f .

$\text{Res}(f, z)$: résidu de la fonction f au point z .

$\sup_{i \in I} x_i$: borne supérieure de la famille des réels (x_i) .

$\inf_{i \in I} x_i$: borne inférieure de la famille des réels (x_i) .

– Convergence de suites

Soient u_n et v_n deux suites réelles. On écrit

$$\left. \begin{array}{l} u_n = \mathcal{O}(v_n) \\ u_n = o(v_n) \\ u_n \sim v_n \end{array} \right\} \quad \text{si la suite } \frac{u_n}{v_n} \quad \left\{ \begin{array}{l} \text{reste bornée} \\ \text{tend vers zéro} \\ \text{tend vers un.} \end{array} \right.$$

$X_n \xrightarrow{\mathcal{L}} X$: (X_n) converge en loi vers X .

$X_n \xrightarrow[n \rightarrow \infty]{\text{p.s.}} X$: (X_n) converge presque sûrement vers X .

$X_n \xrightarrow[n \rightarrow \infty]{\text{P}} X$: (X_n) converge en probabilité vers X .

– Lois

$X \sim Y$: X et Y ont même loi.

i.i.d. : indépendantes et identiquement distribuées.

$\mathcal{N}_d(0, \Gamma)$: loi normale centrée à d dimensions, de matrice de covariance Γ .

$\chi^2(d)$: loi de chi-deux à d degrés de liberté.

– Algèbre linéaire

$\text{tr}(A)$: trace de la matrice A .

Pour une matrice A positive, on note :

$\lambda_{\min}(A)$: plus petite valeur propre de A .

$\lambda_{\max}(A)$: plus grande valeur propre de A .

Table des matières

Remerciements	i
Avant-propos	vii
Notations	xi
1 Définition et propriétés stochastiques de la <i>Chaos Game Representation</i>	1
1.1 Définition	1
1.2 Relation entre CGR et comptage de mots	6
1.3 État de l'art	6
1.4 Propriétés stochastiques	9
1.5 Sur le comportement asymptotique de la transformée de Fourier de π	13
2 Applications statistiques de la CGR	21
2.1 Famille de tests asymptotiques	21
2.1.1 Caractérisation de structure	22
2.1.2 Test d'indépendance	24
2.1.3 Test du caractère markovien	27
2.1.4 Test d'adéquation à une loi	27
2.1.5 Partitions et amélioration de la puissance du test	28
2.1.6 Expérimentations numériques	30
2.1.7 Application à des séquences biologiques	34
2.2 Preuves	37
2.2.1 Preuve du théorème 2.1.3	37
2.2.2 Preuve du théorème 2.1.4	41
2.2.3 Preuve du théorème 2.1.8	43
2.3 Signature génomique et arbres taxonomiques	45
2.3.1 Matrices de distances entre espèces	47
2.3.2 Arbres taxonomiques	50
2.4 Logiciels développés	60
3 Représentation d'une séquence d'ADN en arbre quaternaire	65
3.1 Introduction	66
3.2 Construction de l' <i>arbre-CGR</i> et relation avec l'arbre digital de recherche	66
3.2.1 Algorithme de construction	66

3.2.2	Liens avec les arbres des suffixes	70
3.2.3	Occurrences de mots et recouvrements	73
3.2.4	État de l'art sur les arbres digitaux de recherche	74
3.3	Convergence presque sûre de branches critiques	75
3.3.1	Préambule	75
3.3.2	Lemme préliminaire	80
3.3.3	Minoration de la limite inférieure de la longueur des plus courtes branches	83
3.3.4	Majoration de la limite supérieure de la hauteur	85
3.4	Convergence en probabilité de la profondeur d'insertion	91
3.5	Expérimentations numériques	93
3.6	Logiciels développés	96
3.7	Domaine de définition de la fonction génératrice de la variable représentant la première occurrence d'un mot	96
3.7.1	Preuve de l'assertion ii) de la Proposition 3.3.3	96
3.7.2	Preuve de l'assertion i)	98
4	Convergence des moments dans le TLCPS pour les martingales vectorielles	101
4.1	Théorème de la limite centrale presque sûr et état de l'art	102
4.1.1	Cas scalaire	102
4.1.2	Cas vectoriel	103
4.1.3	Applications	105
4.2	Convergence des moments dans le TLCPS pour les martingales vectorielles	106
4.3	Applications statistiques	108
4.3.1	Estimation des moments, erreurs d'estimation et de prédiction	109
4.3.2	Modèles autorégressifs linéaires	110
4.3.3	Processus de branchement avec immigration	111
4.3.4	Un lien entre la CGR et les processus RCA	114
4.3.5	Sur l'estimateur des moindres carrés pondérés	115
4.4	Preuves	118
4.4.1	Preuve du théorème 4.2.1	118
4.4.2	Preuve du corollaire 4.2.4	131
4.4.3	Preuve du théorème 4.2.5	131
4.4.4	Preuve du corollaire 4.3.1	132
4.4.5	Preuve du corollaire 4.3.2	133
	Conclusions et perspectives	135
	Bibliographie	141

Chapitre 1

Définition et propriétés stochastiques de la *Chaos Game Representation*

Les chapitres 1 et 2 sont le prolongement des travaux [17] réalisés en collaboration avec G. Fayolle et J.-M. Lasgouttes.

L'objectif de ce chapitre est de présenter la *Chaos Game Representation* (CGR). Cette méthode itérative est appliquée pour la première fois aux séquences d'ADN par Jeffrey [47]. Nous commençons par définir cet outil et donnons des exemples commentés de constructions. Puis, après un état de l'art de ses utilisations en bioinformatique, nous énonçons ses principales propriétés stochastiques.

Sommaire

1.1	Définition	1
1.2	Relation entre CGR et comptage de mots	6
1.3	État de l'art	6
1.4	Propriétés stochastiques	9
1.5	Sur le comportement asymptotique de la transformée de Fourier de π	13

1.1 Définition

Le développement actuel de la génétique et l'accélération des programmes de séquençage d'organismes biologiques stimulent une recherche très active sur l'analyse de séquences d'ADN. Il en découle des besoins importants de représentation et de stockage de l'ADN, en particulier pour faciliter la reconnaissance de motifs et détecter des similarités locales ou globales (Roy et al. [80]). La *Chaos Game Representation* (CGR) est à la fois une méthode de représentation graphique et un outil de stockage. Cette méthode itérative fut appliquée pour la première fois aux séquences d'ADN par Jeffrey [47].

Commençons par définir l'algorithme de représentation associé à la CGR. On considère un alphabet fini \mathcal{A} , constitué de d lettres. Pour un borélien borné $S \subset \mathbb{R}^q$, où q est un entier positif, on définit la collection de fonctions affines $\{T_u, u \in \mathcal{A}\}$, liées à un facteur de contraction réel ρ avec $0 < \rho < 1$, par

$$T_u(x) \stackrel{\text{def}}{=} \rho(x + \ell_u), \quad u \in \mathcal{A}, \quad x \in S, \quad \ell_u \in \mathbb{R}^q, \quad (1.1)$$

où, pour tout $u \in \mathcal{A}$, $T_u(S) \subset S$ et

$$T_u(S) \cap T_v(S) = \emptyset, \quad \forall (u, v) \in \mathcal{A}^2, \quad u \neq v. \quad (1.2)$$

Définition 1.1.1. Soit $U_n = u_1 \dots u_n$ une suite de lettres de \mathcal{A} . La CGR de la séquence U_n sur l'ensemble S est la suite de points $\{X_0, \dots, X_n\}$, définie par une position initiale arbitraire X_0 et par la relation récurrente

$$X_{n+1} \stackrel{\text{def}}{=} T_{u_{n+1}}(X_n) = \rho(X_n + \ell_{u_{n+1}}), \quad (1.3)$$

ce qui est équivalent à

$$X_n = \rho^n X_0 + \sum_{k=1}^n \rho^{n-k+1} \ell_{u_k}.$$

À partir d'une séquence de symboles dans un alphabet fini, la CGR associe une trajectoire dans un espace continu, en conservant toutes les propriétés statistiques de la séquence.

Par exemple, \mathcal{A} peut être l'alphabet des 4 nucléotides pour les séquences d'ADN ou l'alphabet des 20 acides aminés pour les protéines. Pour l'ADN, les séquences sont composées de 4 lettres A (adénine), C (cytosine), G (guanine) et T (thymine). La définition de Jeffrey est le cas particulier de la CGR obtenue en choisissant $S = [0, 1]^2$, $\rho = 1/2$. De plus, les 4 lettres sont situées aux quatre sommets du carré unité, avec

$$\ell_A = (0, 0), \quad \ell_C = (0, 1), \quad \ell_G = (1, 1), \quad \ell_T = (1, 0).$$

La relation (1.3) s'écrit alors

$$X_{n+1} = \frac{1}{2}(X_n + \ell_{u_{n+1}})$$

avec $X_0 = (\frac{1}{2}, \frac{1}{2})$. Géométriquement, les nucléotides sont placés de telle sorte que les côtés horizontaux indiquent la composition en bases complémentaires, tandis que les diagonales représentent la composition en purine (A,G) et pyrimidine (C,T). Avec ce choix d'emplacement de lettres, un brin d'ADN et son complémentaire ont des représentations CGR symétriques par rapport à l'axe de symétrie vertical du carré.

On construit la représentation de la façon suivante. Le premier point X_0 est placé au centre du carré. Puis, itérativement, le point X_{n+1} est placé au milieu du segment joignant X_n et le sommet correspondant à la lettre u_{n+1} . La Figure 1.1 illustre la construction de la CGR pour le mot *ATGCGAGTGT*. On peut visualiser sur la Figure 1.2 deux exemples de CGR de séquences d'ADN de longueur 70 000.

La CGR peut aussi être définie sur d'autres ensembles que le carré, par exemple le segment unité $[0, 1]$. Dans le cas général d'un alphabet de d lettres, il est équivalent de prendre l'alphabet $\mathcal{A} = \{0, \dots, d-1\}$ en notant qu'il faut de toute façon ordonner les

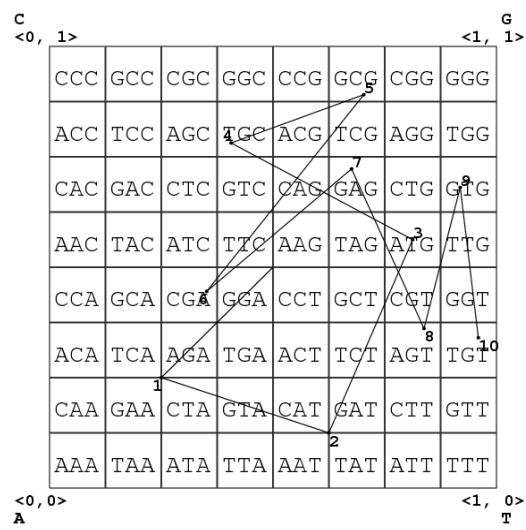


FIG. 1.1: Chaos Game Representation des 10 premiers nucléotides du gène threonine thrA de *E. Coli* : ATGCGAGTGT. Les coordonnées de chaque nucléotide sont calculées récursivement à partir du point initial situé au centre du carré. La séquence est lue de gauche à droite. Le point 3 correspond au premier mot de 3 lettres ATG. Il est situé dans le carré correspondant. Le second mot de 3 lettres TGC correspond au point 4, etc.

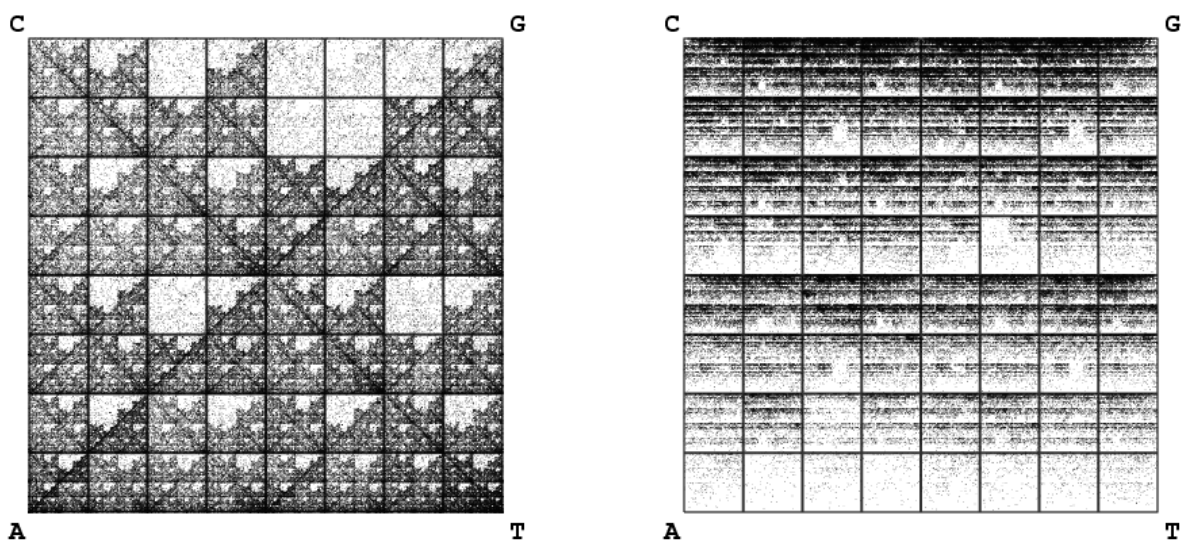


FIG. 1.2: Chaos Game Representation des 70 000 premiers nucléotides du Chromosome 2 d'*Homo Sapiens* à gauche, et de *Streptomyces Coelicolor* sur la droite.

lettres sur le segment. La CGR est alors définie de la façon suivante, avec $\rho = 1/d$ et $\ell_u = u$,

$$X_n = \frac{X_0}{d^n} + \sum_{k=1}^n \frac{u_k}{d^{n-k+1}},$$

où $X_0 = 0$ est le point de départ. On peut voir la construction sur le segment unité du mot *ATGCGAGTGT* donné en exemple sur la Figure 1.3. Les 4 lettres sont placées arbitrairement en

$$\ell_A = 0, \quad \ell_C = 1, \quad \ell_G = 2, \quad \ell_T = 3.$$

Le point initial est placé en X_0 . Puis, itérativement, on construit la suite des points $\{X_0, \dots, X_n\}$ à partir de la relation de récurrence (1.3), en prenant $\rho = \frac{1}{4}$. Cette construction correspond à un développement d -adique, comme nous le précisons dans la section 1.4.

On peut chercher à renforcer la liaison entre les propriétés statistiques, caractérisant la structure de dépendance d'une séquence de lettres de \mathcal{A} , et les propriétés algébriques de l'ensemble des points de la CGR. À cette fin, Gutiérrez et al. [45] proposent une nouvelle CGR pour laquelle le coefficient de contraction ρ n'est plus constant. Ce coefficient dépend de la probabilité d'occurrence de la lettre u_{n+1} . La nouvelle CGR est donnée par la relation récursive

$$X_{n+1} = p_{u_{n+1}} X_n + \ell_{u_{n+1}}, \quad (1.4)$$

où p_v désigne la probabilité de la lettre v , et

$$\ell_A = 0, \quad \ell_C = p_A, \quad \ell_G = p_A + p_C, \quad \ell_T = p_A + p_C + p_G.$$

Avec cette représentation, si $U = U_1 U_2 \dots$ est une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), les U_i n'étant pas nécessairement équiprobables, alors la mesure empirique associée à la suite des points (X_n) de la CGR converge vers la loi uniforme sur $[0, 1[$. Cette représentation ne rentre pas dans le cadre de la Définition 1.1.1. Néanmoins elle possède d'un certain point de vue des propriétés identiques à la CGR classique, comme nous le soulignons dans la section suivante.

La représentation sur le carré, comme celle sur le segment, nécessite de choisir un ordre sur les lettres. Afin de ne pas donner de rôle particulier à certaines lettres de l'alphabet, on peut aussi utiliser la représentation dans un tétraèdre régulier. La première idée naïve, d'affecter un sommet à une lettre u , et de construire la dynamique avec la relation (1.1), a pour conséquence de créer des zones du tétraèdre qui ne seront jamais atteintes. De ce fait, on introduit un bruit visuel avec des images fractales qui n'apportent aucune information sur la séquence. La figure 1.4 montre la construction de la CGR d'une suite de variables aléatoires i.i.d. et équiprobables, avec $\rho = \frac{1}{2}$ et où, pour chaque nucléotide u , $\ell_u = 2\lambda_u$ où λ_u représente les coordonnées du sommet associé à la lettre u . Afin d'établir une bijection entre le tétraèdre et l'ensemble des séquences de 4 lettres, on peut penser à une autre construction. Le point initial X_0 est placé au barycentre du tétraèdre. Puis, itérativement, X_{n+1} est obtenu à partir de X_n de la manière suivante : si l'on note H_{n+1}

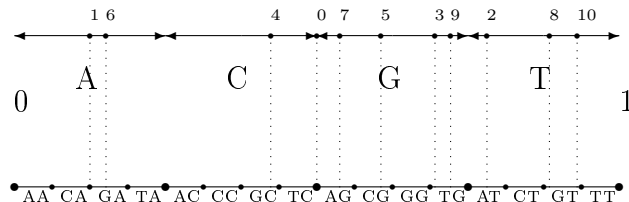


FIG. 1.3: CGR du mot ATGCGAGTGT sur le segment unité. Sur le segment du haut, on a représenté la suite des points. Sur le segment du bas, on a indiqué les ensembles Sw correspondant à tous les dinucléotides w . Cette représentation donne donc la table de fréquences de dinucléotides de la séquence ATGCGAGTGT.

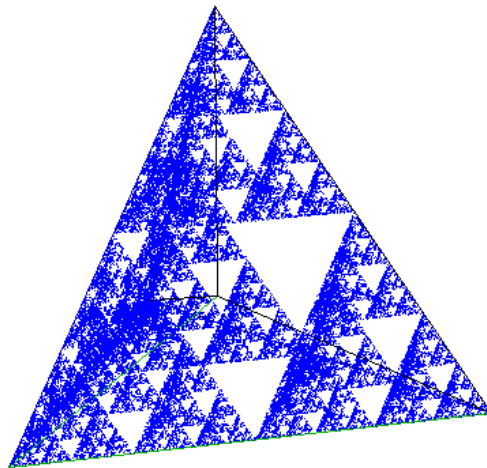


FIG. 1.4: CGR classique sur le tétraèdre de 70 000 nucléotides d'une séquence i.i.d.

le projeté orthogonal de X_n sur la base opposée au sommet correspondant à la lettre u_{n+1} , alors $\overrightarrow{X_n X_{n+1}} = \frac{3}{4} \overrightarrow{X_n H_{n+1}}$. Cette représentation ne rentre pas dans le cadre de la Définition 1.1.1 mais possède certaines propriétés clés de la CGR que nous décrivons dans la section suivante. On peut visualiser sur la Figure 1.5 un exemple de cette CGR sur le tétraèdre pour le Chromosome 2 d'*Homo Sapiens*.

1.2 Relation entre CGR et comptage de mots

On associe au mot $w = u_1 \dots u_n$ l'ensemble Sw défini par

$$Sw \stackrel{\text{def}}{=} \sum_{k=1}^n \rho^{n-k+1} \ell_{u_k} + \rho^n S, \quad (1.5)$$

comme l'illustre la Figure 1.6 pour $S = [0, 1]^2$. Il est équivalent de compter le nombre de points dans le carré Sw ou de compter le nombre d'occurrences du mot w dans la séquence. En effet, Sw contient tous les mots qui ont pour suffixe w (voir la Figure 1.1 pour une matrice de fréquences des trinuécléotides). La CGR est donc une généralisation des tableaux de fréquences de mots (Goldman [40]). Lorsque la CGR est appliquée sur de longues séquences, on peut produire des images dans lesquelles l'intensité de chaque carré Sw est une fonction croissante de la fréquence d'apparition du mot w .

L'une des propriétés importantes de la CGR est que chaque point X_n de la représentation contient toute l'histoire de la séquence X_1, \dots, X_n . En effet, on peut tout d'abord remarquer que l'équation (1.3) entraîne, par construction, que $X_n \in Su_n$. Ainsi, on peut retrouver u_n à partir de X_n et à partir de la relation de récurrence (1.3) : $X_{n-1} = \frac{X_n - \ell_{u_n}}{\rho}$. X_{n-1} est donc déterminé à partir de X_n , et on itère ainsi jusqu'au point initial X_0 .

La représentation de Gutiérrez et al. [45] sur le segment avec coefficient de contraction dépendant de la lettre lue, tout comme la construction dans le tétraèdre à partir des projetés orthogonaux, possèdent ces deux propriétés comme l'illustre la Figure 1.7.

Les deux premiers chapitres de cette thèse abordent les principales propriétés mathématiques de la CGR ainsi que ses applications. Le but est de montrer que la CGR fournit plus d'information sur la distribution d'une séquence que les méthodes classiques liées au comptage de mots.

1.3 État de l'art

À partir d'une subdivision du carré en k quadrants, Almeida et al. [3] définissent une distance, dépendante des corrélations entre les fréquences de points dans chaque quadrant. Cette distance leur permet de faire des comparaisons de gènes et de construire des arbres taxonomiques en s'appuyant notamment sur des techniques d'analyse en composantes principales. Si le nombre de quadrants n'est pas de la forme 4^n , les points de la CGR définissent des fréquences d'oligonucléotides de longueur « fractionnaire ». Les

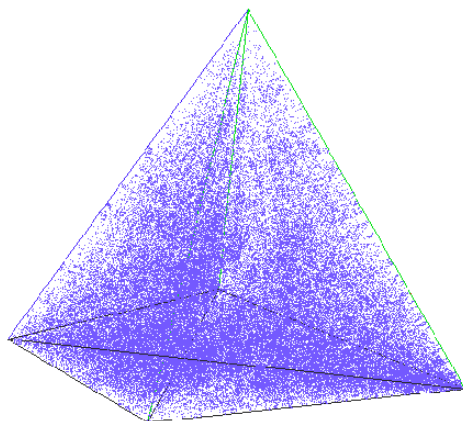


FIG. 1.5: Nouvelle CGR sur le tétraèdre des 70 000 premiers nucléotides du Chromosome 2 d'*Homo Sapiens*.

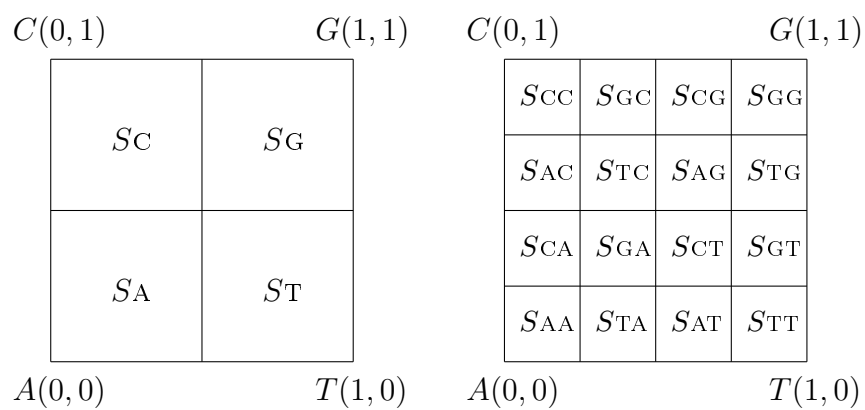


FIG. 1.6: Définitions des carrés correspondant aux nucléotides (à gauche) et aux dinucléotides (à droite) pour la CGR sur $[0, 1]^2$.

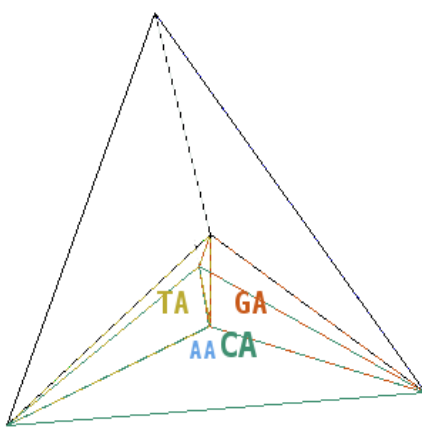
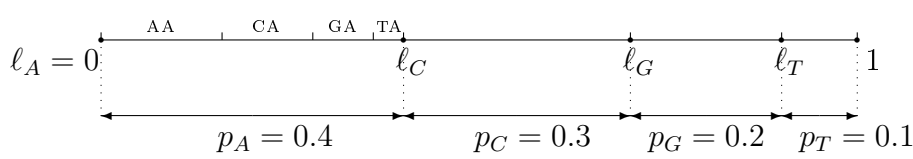


FIG. 1.7: Propriétés autosimilaires des nouvelles CGR définies sur le segment et dans le tétraèdre, où les fonctions itératives ne sont pas affines avec facteur de contraction constant. Les zones Sw sont représentées pour les mots de 2 lettres finissant par A.

auteurs affirment que ces fréquences, de résolution non entière, sont significatives pour les séquences génomiques, à cause de la redondance du code génétique.

De plus, la construction de la CGR est telle que deux séquences se terminant par le même mot de longueur ℓ ont leurs dernières coordonnées contenues dans un même carré de côté $1/2^\ell$. Cette observation se généralise pour mesurer une similarité entre deux séquences, qui correspond à la longueur maximale de mots communs aux deux séquences. Cette longueur est indépendante du choix d'ordonnement des nucléotides sur les sommets du carré. Elle permet à Almeida et al. [3] de construire un ensemble d'algorithmes pour aligner des séquences, avantageux du point de vue du temps de calcul par rapport aux méthodes de scores conventionnelles.

Anh et al. [4] représentent un échantillon de séquences sur des sous-intervalles afin d'obtenir un histogramme de séquences du génome entier. Cet histogramme est appelé *représentation en mesure* du génome. Elle permet de caractériser des séquences d'ADN à partir de la forme de la densité et de l'exposant de leur cascade dans l'analyse multi-fractale.

1.4 Propriétés stochastiques

Dans cette section, on présente les principales propriétés stochastiques de la CGR. Tout d'abord, il est clair que la suite de points définissant la CGR forme une chaîne de Markov d'ordre 1, quel que soit le niveau de dépendance dans la séquence aléatoire $U = u_1 u_2 \dots$ d'éléments de \mathcal{A} . En effet, on a déjà montré que tout le passé de la séquence à un instant donné n peut être retrouvé à partir de la valeur de X_n . Ainsi, les tribus $\sigma(X_n)$ et $\sigma(X_0, (u_k)_{1 \leq k \leq n})$ sont égales. Pour tout $k \leq n$, X_k est donc $\sigma(X_n)$ -mesurable. On a finalement

$$\sigma(X_n) = \sigma(X_1, \dots, X_n).$$

Conditionner par rapport au point X_n revient donc à conditionner par rapport à tout le passé. Cependant, le générateur de la chaîne de Markov (X_n) n'est pas très explicite.

Pour faciliter les notations et la compréhension du processus, on se place dans le cas de la CGR sur le segment $S = [0, 1[$. Les propriétés qui vont suivre sont également vraies dans le cadre de la Définition 1.1.1.

On étudie la relation de récurrence

$$X_n = d^{-1}(X_{n-1} + u_n), \quad n \geq 1, \quad (1.6)$$

où $X_0 \in [0, 1[$ est une constante arbitraire, $d > 1$ un entier positif et $(u_n)_{n \in \mathbb{Z}}$ est une suite stationnaire ergodique à valeurs dans l'alphabet $\mathcal{A} = \{0, 1, \dots, d-1\}$, avec des probabilités d'occurrence des lettres strictement positives. L'équation d'évolution (1.6) est équivalente à

$$X_n = \sum_{k=1}^n \frac{u_k}{d^{n-k+1}} + \frac{X_0}{d^n}. \quad (1.7)$$

Or, on peut voir (1.7) comme le développement en base d de $X_n - d^{-n}X_0$ (dans le sens inverse puisque u_n apparaît en première position). On considère alors la variable aléatoire

$$Y_n = \sum_{k=0}^{\infty} \frac{u_{n-k}}{d^{k+1}}.$$

Il est clair que Y_n est également stationnaire ergodique et vérifie (1.7). De plus, comme les variables u_n sont uniformément bornées, on obtient immédiatement

$$|X_n - Y_n| = \mathcal{O}(d^{-n}),$$

donc

$$\lim_{n \rightarrow \infty} |X_n - Y_n| = 0, \quad p.s.$$

La convergence est géométrique. Soit X la variable aléatoire limite associée à (X_n) et soit π la distribution de X . Pour décider si un point donné dans le support de π correspond à une séquence finie, par exemple $(u_n, u_{n-1}, \dots, u_1)$, ou à une séquence infinie de la forme (u_n, u, u, \dots) contenant la même lettre à partir d'une certaine position, il suffit de placer le point initial à l'un des points fixes de la substitution linéaire σ_u définie par $\sigma_u(x) = d^{-1}(x + u)$, c'est-à-dire à l'un des points

$$a_u = \frac{u}{d-1} \in [0, 1[\quad \text{avec} \quad 0 \leq u \leq d-1.$$

Caractériser la loi limite π de X est une tâche plus difficile. Dans le cas classique où les variables (u_n) sont indépendantes et identiquement distribuées, avec $p_k = \mathbb{P}(u_n = k)$, le processus vérifie les propriétés suivantes (voir par exemple Billingsley [10] et Falconer [34]).

(a) Si u_n est uniforme sur $\{0, 1, \dots, d-1\}$, alors π est simplement la mesure de Lebesgue sur $[0, 1[$. Sinon, π est continue, singulière par rapport à la mesure de Lebesgue et on a la loi forte des grands nombres

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{u_j=k\}} = p_k, \quad \pi \text{ p.s.}$$

(b) Le support S_π de π a la dimension de Hausdorff donnée par

$$\dim_H(S_\pi) = -\frac{1}{\log d} \sum_{i=0}^{d-1} p_i \log p_i. \quad (1.8)$$

Par conséquent, la fonction de répartition $F(x) \stackrel{\text{def}}{=} \mathbb{P}(X \leq x)$ satisfait la condition de Lipschitz d'ordre $\alpha \stackrel{\text{def}}{=} \dim_H(S_\pi)$ (voir Billingsley [10], Falconer [34]), i.e. pour tout $x, h \in [0, 1]$,

$$|F(x+h) - F(x)| = O(h^\alpha).$$

Plusieurs des propriétés ci-dessus sont encore vérifiées sous l'hypothèse que la séquence (u_n) est seulement stationnaire ergodique (SE) .

Démonstration

(a) *Loi des grands nombres.* Elle peut être vue comme une conséquence de l'ergodicité géométrique démontrée ci-dessus (voir Loève [65]). Une autre approche est de remarquer, à partir de l'hypothèse (SE) , qu'il existe un intervalle d -adique

$$I_m = \left[\sum_{k=1}^m \frac{u_k}{d^{m-k+1}}, \sum_{k=1}^m \frac{u_k}{d^{m-k+1}} + \frac{1}{d^m} \right]$$

qui joue le rôle d'*atome récurrent*. En particulier (X_n) est une chaîne de Markov ϕ -irréductible, selon les définitions données par Meyn et Tweedie [69]. Finalement la loi des grands nombres est vérifiée, pour tout borélien B de mesure $\pi(B) > 0$. Ce résultat sera utilisé dans le chapitre 2 pour la convergence des mesures empiriques

$$\hat{\pi}_n(B) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{X_j \in B\}},$$

pour un borélien quelconque $B \in \mathcal{B}(S)$.

(b) *Dimension de Hausdorff.* Avant de calculer la dimension, rappelons très succinctement sa définition. Pour un ensemble F , on note

$$\mathcal{H}_\delta^s(F) = \inf \left\{ \sum_{n=1}^{+\infty} |O_n|^s \mid \{O_n\} \text{ est un } \delta\text{-recouvrement de } F \right\}.$$

On pose également

$$\mathcal{H}^s(F) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(F).$$

La dimension de Hausdorff de l'ensemble F est définie par

$$\dim_H F = \inf \{s \mid \mathcal{H}^s(F) = 0\} = \sup \{s \mid \mathcal{H}^s(F) = \infty\},$$

de sorte que

$$\mathcal{H}^s(F) = \begin{cases} \infty & \text{si } s < \dim_H F, \\ 0 & \text{si } s > \dim_H F. \end{cases}$$

Pour tout nombre $x \in S_\pi$, on note $Sw(x)$ l'*intervalle élémentaire* de longueur $d^{-|w|}$ auquel x appartient. L'argument qui permet d'établir l'égalité (1.8) repose sur la proposition classique brièvement rappelée maintenant.

Proposition 1.4.1 (cf. par exemple Falconer [34]).

$$\begin{cases} \forall x \in S_\pi \limsup_{|w| \rightarrow \infty} \pi(Sw(x))/d^{-s|w|} < c \implies \mathcal{H}^s(S_\pi) \geq 1/c \\ \forall x \in S_\pi \limsup_{|w| \rightarrow \infty} \pi(Sw(x))/d^{-s|w|} > c \implies \mathcal{H}^s(S_\pi) \leq 2^s/c \end{cases}.$$

Dans le cas où la séquence est une suite de réalisations i.i.d., la probabilité de l'événement $\{y \in Sw(x)\}$ est donnée par

$$\pi(Sw(x)) = p_0^{n_0(w)} \dots p_{d-1}^{n_{d-1}(w)},$$

où $n_i(w)$ désigne le nombre d'occurrences de i dans le mot w . D'après la loi des grands nombres, on a

$$\lim_{|w| \rightarrow \infty} \frac{n_i(w)}{|w|} = p_i$$

et par conséquent

$$\begin{aligned} \lim_{|w| \rightarrow \infty} \frac{1}{|w|} \log \frac{\pi(Sw(x))}{d^{-s|w|}} &= \lim_{|w| \rightarrow \infty} \frac{1}{|w|} \sum_{i=0}^{d-1} n_i(w) \log p_i - \frac{1}{|w|} \log d^{-|w|s} \\ &= \sum_{i=0}^{d-1} p_i \log p_i + s \log d. \end{aligned}$$

En posant

$$\alpha \stackrel{\text{def}}{=} -\frac{1}{\log d} \sum_{i=0}^{d-1} p_i \log p_i,$$

on obtient finalement

$$\lim_{|w| \rightarrow \infty} \frac{\pi(Sw(x))}{d^{-s|w|}} = \begin{cases} \infty & \text{si } s > \alpha, \\ 0 & \text{si } s < \alpha. \end{cases}$$

On déduit de la proposition 1.4.1 que $\mathcal{H}^s(S_\pi) = \infty$ si $s < \alpha$ et $\mathcal{H}^s(S_\pi) = 0$ si $s > \alpha$, ce qui conclut la preuve de l'égalité (1.8).

Dans un cadre plus général que l'indépendance, l'argument qui permet de calculer la dimension de Hausdorff repose essentiellement sur le théorème ergodique appliqué à (u_n) . Pour une séquence $w = (w_1 \dots w_n) \in \mathcal{A}^n$, on calcule

$$\mathcal{I} \stackrel{\text{def}}{=} \lim_{|w| \rightarrow \infty} \frac{1}{|w|} \log \pi(Sw(x)) = \lim_{n \rightarrow \infty} \frac{1}{n} \pi(u_n = w_n, \dots, u_1 = w_1). \quad (1.9)$$

Si (u_n) vérifie la condition (SE) , sans autre contrainte, il y a peu d'espoir de simplifier la formule (1.9). Par contre, dans le cas où (u_n) est une chaîne de Markov d'ordre 1 et ergodique, de matrice de transition à coefficients tous strictement positifs, on peut expliciter $\dim_H S_\pi$. Considérons les paires consécutives $u_n u_{n-1}, \dots, u_2 u_1$: elles forment

elles aussi une chaîne de Markov ergodique à valeurs dans \mathcal{A}^2 , de matrice de transition $\mathbf{R} \stackrel{\text{def}}{=} [r(i, j)]$ et de mesure invariante ζ . Alors (1.9) s'écrit

$$\mathcal{I} = \sum_{(i,j) \in \mathcal{A}^2} \zeta(i, j) \log r(i, j). \quad (1.10)$$

On conclut à nouveau grâce à la proposition 1.4.1 pour obtenir la dimension de Hausdorff suivante :

$$\beta = -\frac{1}{\log d} \sum_{(i,j) \in \mathcal{A}^2} \zeta(i, j) \log r(i, j).$$

■

1.5 Sur le comportement asymptotique de la transformée de Fourier de π

Nous considérons le cas où $U = u_1 u_2 \dots$ est une suite de réalisations indépendantes d'une variable aléatoire, à valeurs dans un alphabet fini $\mathcal{A} = \{0, \dots, d-1\}$. L'équation d'évolution

$$X_{n+1} = \rho(X_n + u_{n+1}), \quad \text{avec } \rho = \frac{1}{d},$$

définit une chaîne de Markov (X_n) et on note $\mathbb{F} \stackrel{\text{def}}{=} \{\mathcal{F}_n, n \geq 0\}$ la tribu naturelle associée au processus (X_n) . Il a été démontré dans la section 1.4 que X_n converge en loi vers une variable aléatoire X , prenant ses valeurs sur le segment $[0, 1]$.

Par des méthodes analytiques, nous donnons dans cette section quelques propriétés spectrales (très partielles) satisfaites par la distribution π de X . Notamment, nous obtenons l'exposant de Hölder de π au point $x = 0$, en accord avec les résultats de la section 1.4.

Lemme 1.5.1. Soit $\Phi(t) = \mathbb{E}[e^{itX}]$ la fonction caractéristique de X , définie pour $t \in \mathbb{R}$.

$$\Phi(t) = \prod_{n=0}^{+\infty} g(\rho^n t), \quad \text{avec } g(t) \stackrel{\text{def}}{=} \sum_{v \in \mathcal{A}} p_v e^{itv} \quad p_v \stackrel{\text{def}}{=} \mathbb{P}(U_1 = v). \quad (1.11)$$

Démonstration La démonstration est immédiate à partir de (1.3). On a

$$\begin{aligned} \mathbb{E}\left[e^{itX_{n+1}} \mid \mathcal{F}_n\right] &= \sum_{v \in \mathcal{A}} e^{it\rho(X_n+v)} \mathbb{E}\left[\mathbb{1}_{\{u_{n+1}=v\}} \mid \mathcal{F}_n\right] \\ &= e^{it\rho X_n} \sum_{v \in \mathcal{A}} e^{it\rho v} \mathbb{P}(u_{n+1} = v \mid \mathcal{F}_n). \end{aligned}$$

En notant $\Phi_n(t) \stackrel{\text{def}}{=} \mathbb{E}[e^{itX_n}]$ et en utilisant l'indépendance des variables $\{u_n, n \geq 1\}$, il vient

$$\Phi_{n+1}(t) = g(t) \Phi_n(\rho t) \quad \text{d'où} \quad \Phi_{n+1}(t) = \Phi_0(\rho^{n+1}t) \prod_{k=0}^n g(\rho^k t). \quad (1.12)$$

Lorsque $n \rightarrow \infty$, la suite de fonctions $\Phi_n(t)$ converge vers $\Phi(t)$ définie dans (1.11). De fait, la limite du produit dans (1.12) existe grâce à la convergence de la série

$$\sum_{n=0}^{+\infty} \left| \left(\sum_{v \in \mathcal{A}} p_v e^{itv\rho^n} \right) - 1 \right|.$$

■

Soit le polynôme de degré $d - 1$

$$q(x) \stackrel{\text{def}}{=} \sum_{0 \leq i \leq d-1} p_i x^i.$$

Alors $g(t) = q(e^{i\rho t})$ est un polynôme trigonométrique. Comme q est formellement le produit de $d - 1$ polynômes de degré 1, le comportement asymptotique de $\Phi(t)$ dépend de produits infinis canoniques de la forme

$$P(t) = \prod_{j=0}^{+\infty} \left(1 - a + a e^{i\rho^{j+1}t} \right), \quad (1.13)$$

où a est un nombre complexe tel que $\Re(a) \leq \frac{1}{2}$, car, au prix d'une permutation de a avec $1 - a$, on peut toujours supposer $\left| \frac{a}{1-a} \right| \leq 1$.

Notation Pour toute fonction localement méromorphe $h : z \rightarrow h(z)$, on note $\text{Res}(h, z)$ le résidu de h au point z . On utilisera la fonction spéciale, dite fonction de Lerch (voir par exemple Gradshteyn et Ryzhik [43]),

$$\hat{\Phi}(z, s, v) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} \frac{z^{k+1}}{(v+k)^s},$$

définie pour $-v \notin \mathbb{N}$ et $|z| < 1$ ou $|z| = 1$ et $\Re(s) > 1$. Enfin, on rappelle la définition de la transformée de Mellin $f^*(s)$, $s \in \mathbb{C}$, d'une fonction à valeurs complexe $f(x)$, $x \in \mathbb{R}$.

$$\mathcal{M}[f(x); s] \stackrel{\text{def}}{=} f^*(s) \stackrel{\text{def}}{=} \int_0^{+\infty} f(x) x^{s-1} dx.$$

Théorème 1.5.2. *Pour tous réels $\sigma_1 > 0$ et $t > 0$, le produit $P(t)$ satisfait l'équation*

$$\log P(t) = \text{Res}(F_t, 0) + \sum_{k \in \mathbb{Z}^*} \text{Res}(F_t, \theta_k) + \frac{1}{2i\pi} \int_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} F_t(\theta) d\theta, \quad (1.14)$$

avec

$$F_t(\theta) \stackrel{\text{def}}{=} \frac{1}{\rho^\theta - 1} \Gamma(\theta) \hat{\Phi}\left(\frac{-a}{1-a}, \theta + 1, 1\right) e^{\frac{i\theta\pi}{2}} t^{-\theta}.$$

En outre, au voisinage de $t = +\infty$, on a les comportements asymptotiques

$$\log P(t) = \frac{\log(1-a)}{\log d} \log t + \mathcal{O}(1), \quad (1.15)$$

$$\log \Phi(t) = \frac{\log p_0}{\log d} \log(t) + \mathcal{O}(1). \quad (1.16)$$

L'exposant de Hölder de π en 0 vaut alors

$$\alpha = -\frac{\log p_0}{\log d} > 0. \quad (1.17)$$

Démonstration Il est facile de voir que l'équation (1.16) est une conséquence immédiate de (1.15), en utilisant implicitement la décomposition du polynôme $q(x)$. La preuve de (1.15) repose sur le lemme suivant.

Lemme 1.5.3. *Le produit infini $P(t)$ satisfait, $\forall \sigma$ tel que $-1 < \sigma < 0$, l'équation*

$$\log P(t) = \frac{1}{2i\pi} \int_{\sigma - i\infty}^{\sigma + i\infty} \frac{1}{\rho^\theta - 1} \Gamma(\theta) \hat{\Phi}\left(\frac{-a}{1-a}, \theta + 1, 1\right) e^{\frac{i\theta\pi}{2}} t^{-\theta} d\theta. \quad (1.18)$$

Démonstration On établit, pour $-1 < \theta < 0$ et $\left|\frac{a}{1-a}\right| < 1$, l'identité

$$\begin{aligned} (\log P)^*(\theta) &= \int_0^{+\infty} \sum_{j=0}^{+\infty} \log\left(1 - a + a \exp\left(\frac{it}{4^{j+1}}\right)\right) t^{\theta-1} dt \\ &= \frac{1}{\rho^\theta - 1} \int_0^{+\infty} \log(1 - a + ae^{iu}) u^{\theta-1} du. \end{aligned}$$

Ensuite, on étudie la quantité générique

$$\begin{aligned} k^*(\theta) &\stackrel{\text{def}}{=} \int_0^{+\infty} \log(1 - a + ae^{it}) t^{\theta-1} dt \\ &= -\frac{ia}{\theta} \int_0^{+\infty} \frac{e^{it} t^\theta}{1 - a + ae^{it}} dt, \end{aligned}$$

la seconde égalité ci-dessus résultant d'une intégration par partie. En décomposant la fonction $(1-u)^{-1}$ en série, il vient alors

$$\begin{aligned} k^*(\theta) &= \frac{ia}{\theta} \int_0^{+\infty} \frac{e^{it\theta}}{1-a} \sum_{n=0}^{+\infty} \left(\frac{-a}{1-a}\right)^n e^{int} dt \\ &= -\frac{i}{\theta} \sum_{n=0}^{+\infty} \frac{1}{(n+1)^{\theta+1}} \left(\frac{-a}{1-a}\right)^{n+1} \int_0^{+\infty} e^{iu} u^\theta du. \end{aligned}$$

Finalement,

$$k^*(\theta) = -\frac{i}{\theta} \hat{\Phi}\left(\frac{-a}{1-a}, \theta+1, 1\right) \int_0^{+\infty} e^{iu} u^\theta du.$$

D'autre part, l'égalité « classique » (voir par exemple Dieudonné [26])

$$\int_0^{+\infty} e^{it\theta} dt = e^{\frac{(\theta+1)i\pi}{2}} \Gamma(\theta+1) \quad \forall \theta, \quad -1 < \theta < 0$$

donne

$$k^*(\theta) = \hat{\Phi}\left(\frac{-a}{1-a}, \theta+1, 1\right) e^{\frac{i\theta\pi}{2}} \Gamma(\theta). \quad (1.19)$$

Le théorème de prolongement analytique nous permet d'étendre l'égalité (1.19) à la région $-1 < \Re(\theta) < 0$.

En posant $h^*(\theta) = \frac{k^*(\theta)}{\rho^{\theta-1}}$, il découle de la formule d'inversion de Mellin

$$\log P(t) = \frac{1}{2i\pi} \int_{\sigma-i\infty}^{\sigma+i\infty} h^*(\theta) t^{-\theta} d\theta,$$

ce qui conclut la preuve du lemme 1.5.3. ■

La suite de la preuve du théorème 1.5.2 demande d'intégrer $F_t(\theta) = h^*(\theta)t^{-\theta}$ sur le contour tracé dans la figure 1.8. Ainsi,

$$\frac{1}{2i\pi} \int_{\sigma-i\infty}^{\sigma+i\infty} F_t(\theta) d\theta = -\sum_{k \in \mathbb{Z}} \text{Res}(F_t, \theta_k) + \frac{1}{2i\pi} \int_{\sigma_1-i\infty}^{\sigma_1+i\infty} F_t(\theta) d\theta,$$

avec $\theta_k = -\frac{2ik\pi}{\log \rho}$. On obtient maintenant (1.14) par application du lemme de Jordan qui dit

$$\lim_{R \rightarrow +\infty} \int_{\gamma_3} F_t(\gamma_1(\theta)) \gamma_1'(\theta) d\theta = \lim_{R \rightarrow +\infty} \int_{\gamma_4} F_t(\gamma_2(\theta)) \gamma_2'(\theta) d\theta = 0.$$

Nous poursuivons en analysant chaque terme du membre droit de (1.14). Pour les pôles simples $\theta_k = -\frac{2ik\pi}{\log \rho}$, $k \in \mathbb{Z}^*$, l'expression des résidus est donnée par

$$\text{Res}(F_t, \theta_k) = \frac{\hat{\Phi}\left(\frac{-a}{1-a}, \theta_k+1, 1\right) e^{\frac{i\theta_k\pi}{2}} \Gamma(\theta_k) (t\rho)^{-\theta_k}}{\log \rho}.$$

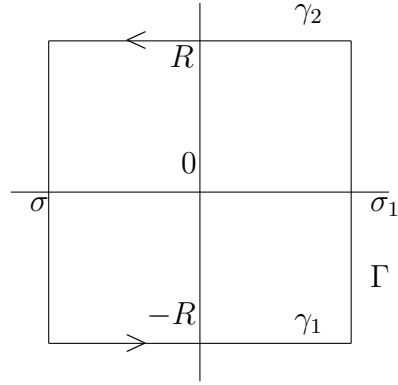


FIG. 1.8: Chemin d'intégration Γ .

Le point 0 est un pôle double associé au résidu

$$\begin{aligned} \text{Res}(F_t, 0) &= \frac{\log(1-a)}{\log \rho} \log t + \log(1-a) \left(\frac{\gamma}{\log \rho} + 1 - \frac{i\pi}{2 \log \rho} \right) \\ &+ \frac{1}{\log \rho} \sum_{k=0}^{+\infty} \left(\frac{-a}{1-a} \right)^{k+1} \frac{\log(k+1)}{k+1}. \end{aligned}$$

De plus, $\forall \sigma_1 > 0$, on a

$$\frac{1}{2i\pi} \int_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} F_t(\theta) d\theta = o(t^{-\sigma_1}).$$

Par conséquent, l'égalité (1.14) se réécrit

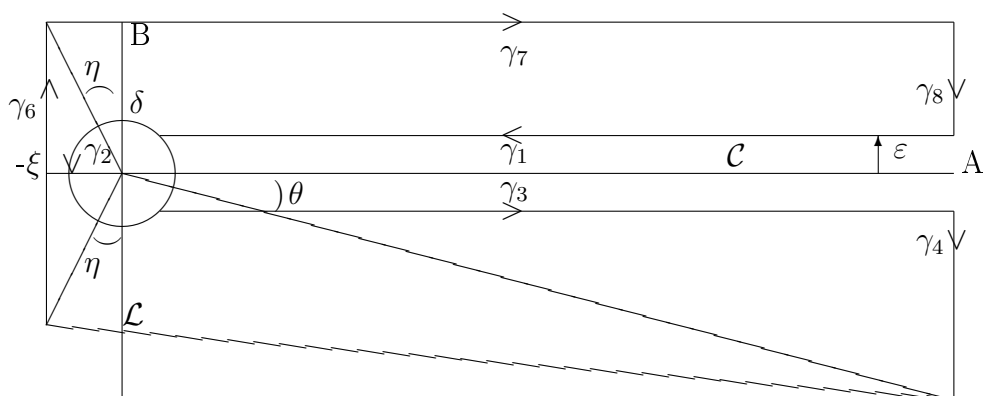
$$\begin{aligned} -\log P(t) &= \frac{\log(1-a)}{\log \rho} \log t + \log(1-a) \left(\frac{\gamma}{\log \rho} + 1 - \frac{i\pi}{2 \log \rho} \right) \\ &+ \frac{1}{\log \rho} \sum_{k=0}^{+\infty} \left(\frac{-a}{1-a} \right)^{k+1} \frac{\log(k+1)}{k+1} \\ &+ \sum_{k \in \mathbb{Z}^*} \text{Res}(F_t, \theta_k) + o(t^{-\sigma_1}). \end{aligned}$$

Il reste à étudier le comportement asymptotique de la série oscillante

$$\sum_{k \in \mathbb{Z}^*} e^{\frac{-k\pi}{2}} \Gamma\left(\frac{-2ik\pi}{\log \rho}\right) t^{\frac{2ik\pi}{\log \rho}} \sum_{n=0}^{+\infty} \left(\frac{-a}{1-a}\right)^{n+1} \frac{1}{(n+1)^{\frac{-2ik\pi}{\log \rho} + 1}}. \quad (1.20)$$

La série formée des termes $k \in \mathbb{N}$ converge, car au voisinage de l'infini,

$$\left| \Gamma\left(\frac{-2ik\pi}{\log \rho}\right) \right| \sim e^{\frac{k\pi^2}{\log \rho}} \left| \frac{k}{\log \rho} \right|^{-1/2}.$$

FIG. 1.9: Chemins d'intégrations \mathcal{C} et \mathcal{L} .

Pour la somme prise sur les entiers négatifs, une analyse plus fine est nécessaire. On utilisera à cet effet la forme intégrale de la fonction de Lerch

$$\sum_{n=1}^{+\infty} \left(\frac{a}{a-1}\right)^n \frac{1}{n^{i\theta_k+1}} = \frac{1}{\Gamma(\theta_k+1)} \int_0^{+\infty} \frac{ae^{-t}}{a-1-ae^{-t}} t^{-i\frac{2k\pi}{\log \rho} \log t} dt.$$

L'étude de cette intégrale est l'objet du lemme suivant.

Lemme 1.5.4.

$$\left| \int_0^{+\infty} \frac{e^{-i\alpha \log t}}{e^t - u} dt \right| = \mathcal{O}\left(\frac{e^{-\theta\alpha}}{|\alpha|}\right).$$

Démonstration En intégrant par parties, il vient

$$\int_0^{+\infty} \frac{e^{-i\alpha \log t}}{e^t - u} dt = \int_0^{+\infty} \frac{te^{-i\alpha \log t} e^t}{(1-i\alpha)(e^t - u)^2} dt.$$

Il sera commode de poser

$$f(z) \stackrel{\text{def}}{=} \frac{ze^{-i\alpha \log z} e^z}{(e^z - u)^2}.$$

On a alors

$$\int_0^{+\infty} \frac{e^{-i\alpha \log t}}{e^t - u} dt = \frac{1}{1-i\alpha} \int_0^{+\infty} f(t) dt.$$

Considérons le chemin d'intégration de la Figure 1.9. Le contour \mathcal{C} est la réunion des arcs γ_1 , γ_2 , et γ_3 , \mathcal{L} étant la réunion de γ_4 , γ_5 , γ_6 , γ_7 et γ_8 . On a évidemment

$$\int_{\mathcal{C}} f(z) dz = - \int_{\mathcal{L}} f(z) dz.$$

Alors

$$\lim_{\substack{\epsilon \rightarrow 0 \\ \delta \rightarrow 0 \\ A \rightarrow +\infty}} \int_{\mathcal{C}} f(z) dz = (e^{2\pi\alpha} - 1) \int_0^{+\infty} \frac{te^{-i\alpha \log t} e^t}{(e^t - u)^2} dt,$$

d'où

$$\int_0^{+\infty} \frac{e^{-i\alpha \log t}}{e^t - u} dt = \frac{1}{1 - i\alpha} \frac{1}{e^{2\pi\alpha} - 1} \lim_{\substack{\epsilon \rightarrow 0 \\ \delta \rightarrow 0 \\ A \rightarrow +\infty}} - \int_{\mathcal{L}} f(z) dz.$$

Il reste à étudier le comportement de f sur le contour \mathcal{L} . Notons que

$$\lim_{A \rightarrow +\infty} \left| \int_{\gamma_7} f(z) dz \right| \leq e^{(\frac{\pi}{2} + \eta)\alpha} \int_{-\xi}^{+\infty} \frac{e^t |\gamma_7(t)|}{|e^t - u|^2} dt. \quad (1.21)$$

On a également

$$\lim_{\substack{\epsilon \rightarrow 0 \\ A \rightarrow +\infty}} \left| \int_{\gamma_6} f(z) dz \right| = \lim_{A \rightarrow +\infty} \left| \int_{\gamma_4} f(z) dz \right| = 0.$$

De l'inégalité

$$\left| \int_{\gamma_5} f(z) dz \right| \leq e^{(2\pi - \theta)\alpha} \int_{-\xi}^A \frac{e^t |\gamma_5(t)|}{(e^t - |u|)^2} dt,$$

l'intégrale de droite étant convergente, on déduit l'existence d'une constante $C(u, \xi)$ telle que

$$\lim_{A \rightarrow +\infty} \left| \int_{\gamma_5} f(z) dz \right| \leq C(u, \xi) e^{(2\pi - \theta)\alpha}.$$

La preuve du lemme 1.5.4 est achevée. ■

La dernière étape consiste à faire tendre θ vers $\frac{\pi}{2}$: puisque

$$\left| \sum_{n=1}^{+\infty} \left(\frac{a}{a-1} \right)^n \frac{1}{n^{i\theta_k + 1}} \right| = \mathcal{O}\left(\frac{1}{|k|^{3/2}} \right),$$

les comportements asymptotiques (1.15) et (1.16) sont démontrés.

La valeur de l'exposant de Hölder de π en 0 est une conséquence directe des formules d'inversion de Mellin.

Le cas $|\frac{a}{a-1}| = 1$, i.e. $\Re(a) = \frac{1}{2}$, se traiterait de façon analogue, avec quelques particularités techniques, et c'est pourquoi nous l'omettons. ■

Chapitre 2

Applications statistiques de la *Chaos Game Representation*

Comment utiliser la CGR et l'information qu'elle contient ? La CGR fournit-elle plus d'information que les méthodes de comptage de mots classiques ?

A partir de propriétés sur la nature invariante des points de la CGR, on construit une nouvelle famille de tests caractérisant l'ordre d'une chaîne de Markov homogène. Par ailleurs, on propose une généralisation de la notion d'abondance relative de dinucléotides comme *signature génomique*, qui permet de construire des arbres taxonomiques.

Sommaire

2.1	Famille de tests asymptotiques	21
2.1.1	Caractérisation de structure	22
2.1.2	Test d'indépendance	24
2.1.3	Test du caractère markovien	27
2.1.4	Test d'adéquation à une loi	27
2.1.5	Partitions et amélioration de la puissance du test	28
2.1.6	Expérimentations numériques	30
2.1.7	Application à des séquences biologiques	34
2.2	Preuves	37
2.2.1	Preuve du théorème 2.1.3	37
2.2.2	Preuve du théorème 2.1.4	41
2.2.3	Preuve du théorème 2.1.8	43
2.3	Signature génomique et arbres taxonomiques	45
2.3.1	Matrices de distances entre espèces	47
2.3.2	Arbres taxonomiques	50
2.4	Logiciels développés	60

2.1 Famille de tests asymptotiques

Modéliser une séquence biologique consiste à la considérer comme un objet probabiliste, dépendant de plusieurs paramètres que l'on cherche à estimer. Un modèle markovien d'ordre m donnera par exemple des résultats satisfaisants pour prédire les comptages

de mots de longueur $m + 1$. On pourra également dépister une fonctionnalité biologique si un mot est significativement sur- ou sous-représenté, dans une séquence donnée, par rapport au modèle choisi. On peut consulter Rocha et al. [79] pour l'analyse de palindromes dans les génomes de bactérie ou El Karoui et al. [30] pour l'analyse de Chi. Bien que Churchill [22] et Muri [70] mettent en évidence une hétérogénéité dans les séquences d'ADN à l'aide de modèles de Markov cachés, nous nous limiterons dans ce travail à des séquences stationnaires ergodiques. Dans ce contexte, Reinert et al. [77] proposent un test afin de déterminer un ordre de dépendance approprié dans les séquences, en utilisant la statistique de Pearson.

Nous proposons ici une famille de tests asymptotiques pour déterminer la structure d'une séquence aléatoire de symboles dans un alphabet fini (indépendance, chaîne de Markov d'ordre m , etc). La construction de ces tests utilise la CGR. En particulier, ils reposent sur des statistiques qui dépendent de toute l'histoire de la séquence, contrairement aux tests classiques.

Ces tests peuvent avoir d'autres applications, par exemple en cryptographie. En effet, ils permettent de détecter des déviations par rapport à des séquences indépendantes et donc d'apprécier la qualité de certains générateurs aléatoires ou pseudo-aléatoires (Knuth [55], L'Ecuyer [60], Menezes et al. [68]).

2.1.1 Caractérisation de structure

Il est possible de donner un critère simple permettant de savoir si une CGR a été obtenue à partir d'une séquence i.i.d. ou non. Avant d'énoncer ce critère, introduisons une notation qui sera utilisée dans toute la suite. La correspondance bijective entre l'ensemble des séquences possibles et l'ensemble des points de la CGR suggère de noter, pour un mot $w = v_1 \dots v_m$ constitué de lettres dans un alphabet fini \mathcal{A} , et pour un ensemble $B \subset S$,

$$Bw \stackrel{\text{def}}{=} T_{v_m} \circ \dots \circ T_{v_1}(B). \quad (2.1)$$

Cette définition coïncide clairement avec (1.5) lorsque $B = S$. Un point X_n de la CGR, associé à une suite de lettres $U_n = u_1 \dots u_n$ de \mathcal{A} se trouve dans Bw si $X_{n-m} \in B$ et $u_{n-m+1} \dots u_n = v_1 \dots v_m$.

Supposons maintenant que U soit une séquence stationnaire ergodique, et notons π la mesure limite invariante de sa CGR sur S . La propriété de π énoncée ci-dessous est très utile pour la construction du test.

Proposition 2.1.1. *La séquence aléatoire stationnaire U est indépendante et identiquement distribuée si et seulement si*

$$\pi(Bu) = \pi(B)\pi(Su), \quad \forall u \in \mathcal{A} \quad \forall B \subset S. \quad (2.2)$$

Démonstration Si U est i.i.d. alors l'événement $\{X_{n+1} \in Su\} = \{u_{n+1} = u\}$ est indépendant de l'événement $\{X_n \in B\}$. On déduit alors de (1.3) et (2.1) que

$$\mathbb{P}(X_{n+1} \in Su)\mathbb{P}(X_n \in B) = \mathbb{P}(X_{n+1} \in Su, X_n \in B) = \mathbb{P}(X_{n+1} \in Bu).$$

Réciproquement, supposons que (2.2) soit vérifiée. Soit $v_1 \dots v_m$ une séquence finie arbitraire. En choisissant $u = v_m$ et $B = Sv_1 \dots v_{m-1}$, il est clair que

$$\pi(Sv_1 \dots v_m) = \pi((Sv_1 \dots v_{m-1})v_m) = \pi(Sv_1 \dots v_{m-1})\pi(Sv_m).$$

Il en découle immédiatement par récurrence que

$$\pi(Sv_1 \dots v_m) = \pi(Sv_1)\pi(Sv_2) \cdots \pi(Sv_m).$$

On peut conclure que la séquence aléatoire U est i.i.d. car la suite de lettres $v_1 \dots v_m$ peut être arbitrairement choisie. ■

Il est possible d'étendre cette caractérisation au cas markovien.

Proposition 2.1.2. *La séquence aléatoire stationnaire U est une chaîne de Markov d'ordre m si et seulement si,*

$$\pi(Sw)\pi(Bwu) = \pi(Swu)\pi(Bw), \quad \forall B \subset S, \quad \forall w \in \mathcal{A}^m, \quad \forall u \in \mathcal{A}. \quad (2.3)$$

En particulier le rapport $\pi(Bwu)/\pi(Bw)$ ne dépend pas de B .

Démonstration Si U est une chaîne de Markov d'ordre m et n un entier ≥ 0 , on a

$$\frac{\pi(Bwu)}{\pi(Bw)} = \mathbb{P}(u_{n+1} = u | X_n \in Bw) = \mathbb{P}(u_{n+1} = u | X_n \in Sw) = \frac{\pi(Swu)}{\pi(Sw)},$$

d'où l'on déduit (2.3). Réciproquement, si l'on suppose que (2.3) est vérifiée, on considère une séquence finie arbitraire $v_1 \dots v_{n+1}$ avec $n \geq m$. On choisit $u = v_{n+1}$, $w = v_{n-m+1} \dots v_n$ et $B = Sv_1 \dots v_{n-m}$. Ainsi on a

$$\frac{\pi(Sv_1 \cdots v_{n+1})}{\pi(Sv_1 \cdots v_n)} = \frac{\pi(Bwu)}{\pi(Bw)} = \frac{\pi(Swu)}{\pi(Sw)} = \frac{\pi(Sv_{n-m+1} \cdots v_{n+1})}{\pi(Sv_{n-m+1} \cdots v_n)}.$$

Par conséquent, on a

$$\begin{aligned} \mathbb{P}(u_{n+1} = v_{n+1} | u_1 = v_1, \dots, u_n = v_n) &= \frac{\pi(Sv_1 \cdots v_{n+1})}{\pi(Sv_1 \cdots v_n)} = \frac{\pi(Sv_{n-m+1} \cdots v_{n+1})}{\pi(Sv_{n-m+1} \cdots v_n)} \\ &= \mathbb{P}(u_{n+1} = v_{n+1} | u_{n-m+1} = v_{n-m+1}, \dots, u_n = v_n). \end{aligned}$$

Comme la suite de lettres v_1, \dots, v_n peut être arbitrairement choisie, la séquence aléatoire U est une chaîne de Markov d'ordre au plus m . ■

2.1.2 Test d'indépendance

On note respectivement H_0 , H_m et H les hypothèses suivantes : « $U_n = u_1 \dots u_n$ est une séquence indépendante et identiquement distribuée », « U_n est une chaîne de Markov d'ordre m » et « U_n est une séquence stationnaire ergodique ». On souhaite réaliser un test de l'hypothèse H_0 contre $H \setminus H_0$, et de H_m contre $H \setminus H_m$. On rappelle (cf. Chapitre 1) que la notation $\hat{\pi}_n$ désigne la mesure empirique associée à la CGR (X_n). Pour tout ensemble B , on note donc

$$\hat{\pi}_n(B) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{X_j \in B\}}.$$

La Proposition 2.1.1 suggère de choisir une région de rejet de la forme

$$\left| \hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) \right| > \epsilon$$

avec $\epsilon > 0$. Pour construire un test asymptotique de niveau α , il reste à ajuster et choisir ϵ en utilisant le théorème de la limite centrale suivant.

Théorème 2.1.3. *Soit u_α le $(1 - \frac{\alpha}{2})$ -quantile de la loi normale centrée réduite, c'est-à-dire le réel u_α tel que $P(|Y| > u_\alpha) = \alpha$ où Y suit la loi normale centrée réduite. On définit également*

$$\hat{\sigma}_n(B, v) \stackrel{\text{def}}{=} \sqrt{\left(\hat{\pi}_n(Sv)(1 - \hat{\pi}_n(Sv)) \right) \left(\hat{\pi}_n(B)(1 - \hat{\pi}_n(B)) \right)}.$$

Alors, l'ensemble

$$\left\{ \left| \hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) \right| > u_\alpha \frac{\hat{\sigma}_n(B, v)}{\sqrt{n}} \right\} \quad (2.4)$$

est une région de rejet de niveau asymptotique α de l'hypothèse nulle H_0 contre $H \setminus H_0$. Plus précisément,

$$\lim_{n \rightarrow +\infty} \mathbb{P}_{H_0} \left(\left| \hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) \right| > u_\alpha \frac{\hat{\sigma}_n(B, v)}{\sqrt{n}} \right) \leq \alpha.$$

Démonstration La preuve du théorème 2.1.3 est donnée dans la Section 2.2.1. ■

Pour appliquer le théorème 2.1.3, il faut choisir un nucléotide v et un ensemble $B \subset S$. Un choix *optimal* de B et v dépend de la distribution (p_u) , inconnue en pratique. Pour tester H_0 contre H_m , $m \geq 1$, Reinert et al. [77] proposent d'utiliser la statistique de Pearson

$$X^2 \stackrel{\text{def}}{=} \sum_{u,v \in \mathcal{A}} \frac{\left(N(uv) - N(u \cdot)N(\cdot v)/(n-1) \right)^2}{N(u \cdot)N(\cdot v)/(n-1)}, \quad (2.5)$$

où $N(uv)$ désigne le nombre d'occurrences de uv dans la séquence U_n , $N(u\cdot)$ (resp. $N(\cdot v)$) représente le nombre de mots de 2 lettres commençant par u (resp. terminant par v). Sous H_0 , X^2 suit asymptotiquement une loi de chi-deux à $(d-1)^2$ degrés de liberté.

Dans le cas particulier où l'on choisit $B = Su$, la statistique de test utilisée au théorème 2.1.3 se lit

$$\begin{aligned} & \hat{\pi}_n(Suv) - \hat{\pi}_n(Sv)\hat{\pi}_n(Su) \\ &= \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{Suv}(X_j) - \frac{1}{n^2} \sum_{j=1}^n \mathbb{1}_{Sv}(X_j) \sum_{i=1}^n \mathbb{1}_{Su}(X_i). \end{aligned}$$

Par analogie avec (2.5), on propose un test de H_0 contre $H \setminus H_0$, de région de rejet de la forme

$$\left\{ \sum_{B \in \mathcal{P}, v \in \mathcal{A}} R_n^2(B, v) > q_\alpha \right\}, \quad (2.6)$$

pour une partition \mathcal{P} de S , avec

$$R_n(B, v) \stackrel{\text{def}}{=} \frac{\sqrt{n}(\hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B))}{\sqrt{\hat{\pi}_n(Sv)\hat{\pi}_n(B)}}. \quad (2.7)$$

Pour la partition $\{Su, u \in \mathcal{A}\}$ constituée des carrés associés aux d lettres de l'alphabet, la statistique de test (2.6) est analogue à X^2 . Le choix d'utiliser une partition de S pour la construction du test est motivé par l'analogie avec la statistique de Pearson. En effet, pour le choix d'un seul ensemble $B = Su$, seuls les mots de deux lettres commençant par u sont considérés dans la statistique de test

$$\left\{ \sum_{v \in \mathcal{A}} R_n^2(Su, v) > q_\alpha \right\}.$$

Théorème 2.1.4 (Test d'indépendance). *Pour tout ensemble $B \subset S$,*

$$\frac{1}{1 - \hat{\pi}_n(B)} \sum_{v \in \mathcal{A}} R_n^2(B, v) \stackrel{\mathcal{L}}{\rightsquigarrow} \chi^2(d-1). \quad (2.8)$$

Soit $q_\alpha(d)$ le $(1 - \alpha)$ -quantile de la loi de chi-deux $\chi^2(d)$. Alors, l'ensemble

$$\left\{ \frac{1}{1 - \hat{\pi}_n(B)} \sum_{v \in \mathcal{A}} R_n^2(B, v) > q_\alpha(d-1) \right\} \quad (2.9)$$

est une région de rejet d'un test de niveau asymptotique α de H_0 contre $H \setminus H_0$.

Plus généralement, pour une partition \mathcal{P} de S , avec $|\mathcal{P}| = K > 1$, où $|\mathcal{P}|$ désigne la taille de la partition \mathcal{P} , l'ensemble

$$\left\{ \sum_{B \in \mathcal{P}, v \in \mathcal{A}} R_n^2(B, v) > q_\alpha [(d-1)(K-1)] \right\} \quad (2.10)$$

est une région d'un test asymptotique de niveau α , de H_0 contre $H \setminus H_0$.

Démonstration La preuve du théorème 2.1.4 est donnée dans la Section 2.2.2. ■

Remarque 2.1.5. Un test construit à partir d'une partition de deux ensembles, de région de rejet (2.10), est exactement le même qu'un test construit à partir de l'un de ces deux ensembles, B , de région de rejet (2.9). En effet, soit \bar{B} le complémentaire de B dans S . On a $\hat{\pi}_n(Bv) + \hat{\pi}_n(\bar{B}v) = \hat{\pi}_n(Sv)$. Il en découle que

$$\begin{aligned} & R_n^2(B, v) + R_n^2(\bar{B}, v) \\ &= \frac{n(\hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B))^2}{\hat{\pi}_n(Sv)\hat{\pi}_n(B)} + \frac{n(-\hat{\pi}_n(Bv) + \hat{\pi}_n(Sv)\hat{\pi}_n(B))^2}{\hat{\pi}_n(Sv)(1 - \hat{\pi}_n(B))} \\ &= \frac{n(\hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B))^2}{\hat{\pi}_n(Sv)\hat{\pi}_n(B)(1 - \hat{\pi}_n(B))} \\ &= \frac{1}{1 - \hat{\pi}_n(B)} R_n^2(B, v). \end{aligned}$$

Pour apprécier la qualité des tests définis dans le théorème 2.1.4, on peut calculer la puissance asymptotique, c'est-à-dire la probabilité asymptotique sous $H \setminus H_0$ des ensembles (2.9) et (2.10).

Théorème 2.1.6. *Supposons que l'hypothèse $H \setminus H_0$ soit vérifiée, et soient $B \subset S$ et $v \in \mathcal{A}$ tels que*

$$\pi(Bv) \neq \pi(Sv)\pi(B). \quad (2.11)$$

Alors la puissance asymptotique du test construit à partir de la région de rejet (2.9) est 1, c'est-à-dire que le test est asymptotiquement consistant. Si de plus B est l'un des ensembles formant la partition \mathcal{P} , alors le test construit à partir de la région de rejet (2.10) est asymptotiquement consistant.

Démonstration On déduit de la convergence des mesures empiriques que

$$\lim_{n \rightarrow \infty} \hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) = \pi(Bv) - \pi(Sv)\pi(B) \quad \text{p.s.}$$

Ainsi, sous l'hypothèse (2.11),

$$\lim_{n \rightarrow \infty} \sqrt{n} \left| \hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) \right| = +\infty \quad \text{p.s.}$$

et donc

$$\lim_{n \rightarrow \infty} \sum_{v \in \mathcal{A}} R_n^2(B, v) = +\infty \quad \text{p.s.} \quad (2.12)$$

On en déduit aisément les assertions du théorème 2.1.6. ■

Remarque 2.1.7. Sous $H \setminus H_0$, la convergence (2.12) est d'autant plus rapide que le nombre d'ensembles B de la partition satisfaisant (2.11) augmente. Par conséquent, la puissance du test convergera plus vite vers 1 pour une partition contenant de nombreux ensembles B satisfaisant (2.11). Cependant, il est nécessaire de trouver un compromis car la convergence vers le niveau est d'autant plus lente que le nombre d'ensembles formant la partition augmente.

2.1.3 Test du caractère markovien

La statistique correspondant à (2.7) dans le cas markovien est le rapport

$$R_n(B, w, u) \stackrel{\text{def}}{=} \sqrt{n} \frac{\hat{\pi}_n(Sw)\hat{\pi}_n(Bwu) - \hat{\pi}_n(Swu)\hat{\pi}_n(Bw)}{\sqrt{\hat{\pi}_n(Sw)\hat{\pi}_n(Swu)\hat{\pi}_n(Bw)}}$$

Comme dans le théorème 2.1.4, la statistique R_n est utilisée pour construire un test de H_m contre $H \setminus H_m$.

Théorème 2.1.8 (Test de caractérisation markovienne). (i) Pour une partition \mathcal{P} de S , avec $|\mathcal{P}| = K > 1$, l'ensemble

$$\left\{ \sum_{wu \in \mathcal{A}^m \times \mathcal{A}, B \in \mathcal{P}} R_n^2(B, w, u) > q_\alpha \left[d^m(d-1)(K-1) \right] \right\} \quad (2.13)$$

est une région de rejet d'un test de niveau asymptotique α , de H_m contre $H \setminus H_m$.
(ii) Sous $H \setminus H_m$, et en supposant qu'il existe $B \in \mathcal{P}$ tel que

$$\pi(Sw)\pi(Bwu) \neq \pi(Bw)\pi(Swu), \quad (2.14)$$

le test construit à partir de la région de rejet (2.13) est asymptotiquement consistant.

Démonstration La preuve du théorème 2.1.8 est donnée dans la Section 2.2.3. ■

2.1.4 Test d'adéquation à une loi

Avec la même méthode, on peut construire des tests d'adéquation à une loi, basés sur des statistiques analogues à celles utilisées pour les tests de structure.

Théorème 2.1.9. (i) Pour une partition \mathcal{P} de S , avec $|\mathcal{P}| = K > 1$, et pour une matrice de transition Q donnée, l'ensemble

$$\left\{ \sum_{wu \in \mathcal{A}^m \times \mathcal{A}, B \in \mathcal{P}} n \frac{(\hat{\pi}_n(Bwu) - Q(w, u)\hat{\pi}_n(Bw))^2}{Q(w, u)\hat{\pi}_n(Bw)} > q_\alpha [d^m(d-1)K] \right\}$$

est une région de rejet d'un test de niveau asymptotique α , de l'hypothèse $H'_m \ll U_n$ est une chaîne de Markov d'ordre m et de matrice de transition Q » contre $H \setminus H'_m$.
(ii) Sous l'hypothèse $H \setminus H'_m$, et en supposant qu'il existe $B \in \mathcal{P}$ tel que

$$\pi(Bwu) \neq Q(w, u)\pi(Bw),$$

le test est asymptotiquement consistant.

Démonstration La preuve est exactement identique à celle du théorème 2.1.8. Elle est laissée au lecteur. ■

2.1.5 Partitions et amélioration de la puissance du test

Pour éviter un choix rigide d'une partition unique, on peut proposer une généralisation du test précédent à une collection de partitions de S . L'idée est inspirée de la méthode de Bonferroni décrite dans Baraud et al. [5]. Pour une collection finie $\Pi = \{\mathcal{P}_1, \dots, \mathcal{P}_p\}$, avec $K_j \stackrel{\text{def}}{=} |\mathcal{P}_j|$, H_0 est rejetée dès que l'une des partitions \mathcal{P}_j vérifie

$$\sum_{v \in \mathcal{A}, B \in \mathcal{P}_j} R_n^2(B, v) - q_{\alpha_j} [(d-1)(K_j - 1)] > 0.$$

Il reste à choisir le niveau α_j de chaque partition \mathcal{P}_j pour obtenir un niveau global égal à α pour la collection Π . C'est-à-dire, si l'on note

$$Z \stackrel{\text{def}}{=} \sup_{1 \leq j \leq p} \left\{ \sum_{B \in \mathcal{P}_j, v \in \mathcal{A}} R_n^2(B, v) - q_{\alpha_j} [(d-1)(K_j - 1)] \right\},$$

alors la suite $(\alpha_j)_{1 \leq j \leq p}$ doit être choisie comme un ensemble de réels dans l'intervalle $]0, 1[$ tels que

$$\mathbb{P}_{H_0}(Z > 0) \leq \alpha. \quad (2.15)$$

Par conséquent, $\{Z > 0\}$ est une région de rejet d'un test asymptotique de niveau α de H_0 contre $H \setminus H_0$.

Théorème 2.1.10. Sous l'hypothèse $H \setminus H_0$, on suppose qu'il existe $B \subset S$ et $v \in \mathcal{A}$, tels que (2.11) soit vérifiée. Si $\Pi = \{\mathcal{P}_1, \dots, \mathcal{P}_p\}$ est une collection de partitions de S telle que $B \in \mathcal{P}_j$ pour un entier $j \in \{1, \dots, p\}$, si la suite (α_j) satisfait (2.15), alors le test de région de rejet $\{Z > 0\}$ est asymptotiquement consistant.

Démonstration La preuve repose sur les mêmes arguments que celle du théorème 2.1.6.

■

En pratique, la suite $(\alpha_j, 1 \leq j \leq p)$ est choisie selon l'une des procédures suivantes. Dans la première, on prend

$$\sum_{1 \leq j \leq p} \alpha_j = \alpha.$$

En effet, on a

$$\begin{aligned} \mathbb{P}_{H_0}(Z > 0) &\leq \sum_{1 \leq j \leq p} \mathbb{P}_{H_0} \left(\sum_{\substack{B \in \mathcal{P}_j \\ v \in \mathcal{A}}} R_n^2(B, v) - q_{\alpha_j} [(d-1)(K_j-1)] > 0 \right) \\ &= \sum_{1 \leq j \leq p} \alpha_j = \alpha. \end{aligned}$$

Dans la seconde procédure, on affecte à tous les α_j la même valeur γ , qui est le γ -quantile de la variable aléatoire

$$\inf_{1 \leq j \leq p} Q_j \left(\sum_{B \in \mathcal{P}_j, v \in \mathcal{A}} R_n^2(B, v) \right), \quad (2.16)$$

où Q_j désigne la queue de distribution d'un chi-deux à $(d-1)(K_j-1)$ degrés de liberté. Il advient que l'inégalité (2.15) est bien vérifiée puisque

$$\begin{aligned} \alpha &= \mathbb{P}_{H_0}(Z > 0) \\ &= 1 - \mathbb{P}_{H_0} \left(\forall j \in \{1, \dots, p\} \sum_{B \in \mathcal{P}_j, v \in \mathcal{A}} R_n^2(B, v) \leq q_\gamma [(d-1)(K_j-1)] \right) \\ &= \mathbb{P}_{H_0} \left(\inf_{1 \leq j \leq p} Q_j \left(\sum_{B \in \mathcal{P}_j, v \in \mathcal{A}} R_n^2(B, v) \right) \leq \gamma \right). \end{aligned}$$

Comme dans le cas i.i.d, le test défini dans le théorème 2.1.8 peut être généralisé à une collection de partitions $\Pi = \{\mathcal{P}_1, \dots, \mathcal{P}_p\}$. La statistique correspondant à Z est

$$Z_{(m)} \stackrel{\text{def}}{=} \sup_{1 \leq j \leq p} \left\{ \sum_{B \in \mathcal{P}_j, v \in \mathcal{A}} R_n^2(B, w, v) - q_{\alpha_j} \left[d^m (d-1)(K_j-1) \right] \right\}.$$

De plus, dans la seconde procédure, la statistique (2.16) est remplacée par

$$\inf_{1 \leq k \leq p} Q_j^m \left(\sum_{B \in \mathcal{P}_j, v \in \mathcal{A}} R_n^2(B, w, v) \right),$$

où Q_j^m désigne la queue de distribution d'un Chi-deux à $d^m(d-1)(K_j-1)$ degrés de liberté.

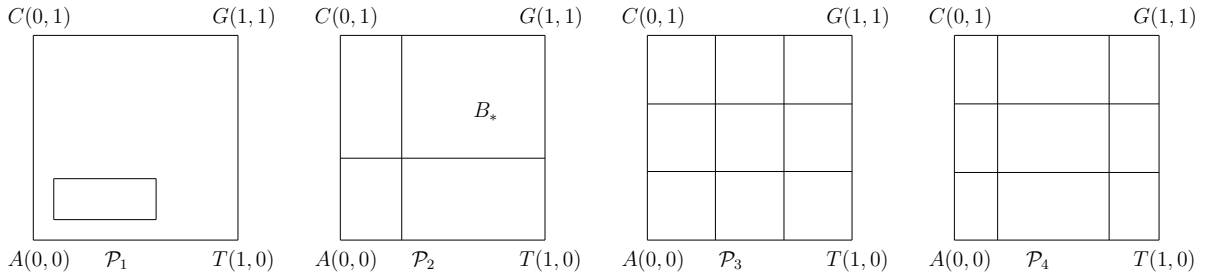


FIG. 2.1: Les quatre différentes partitions du carré unité $[0, 1]^2$ choisies pour le test.

Théorème 2.1.11. *Sous $H \setminus H_m$, on suppose qu'il existe $B \subset S$, $v \in \mathcal{A}$, et $w \in \mathcal{A}^m$ tels que (2.14) soit vérifiée. Si $\Pi = \{\mathcal{P}_1, \dots, \mathcal{P}_p\}$ est une collection de partitions de S telle que $B \in \mathcal{P}_j$ pour un entier $j \in \{1, \dots, p\}$, alors le test est asymptotiquement consistant.*

Démonstration La preuve repose sur les mêmes arguments que celle du théorème 2.1.8.

■

2.1.6 Expérimentations numériques

Dans cette section, on compare numériquement les tests décrits ci-dessus au test de Pearson. La CGR est calculée sur le carré unité, et l'alphabet \mathcal{A} est composé des 4 nucléotides $\{A, C, G, T\}$.

Test d'indépendance.

Pour plusieurs valeurs de n , on génère 1 000 chaînes de Markov de longueur n et d'ordre m , dont les matrices de transition sont choisies uniformément et indépendamment les unes des autres. Afin de mesurer l'influence du choix de la partition sur la performance du test décrit dans le théorème 2.1.4, on choisit arbitrairement plusieurs partitions, représentées sur la Figure 2.1. Les ensembles formant chaque partition sont tels qu'ils ne coïncident pas avec des unions de carrés Sw , quels que soient les mots w . Pour apprécier qualitativement l'information fournie dans une partition plutôt que dans un seul ensemble, on sélectionne dans la première expérience un ensemble $B_* \in \mathcal{P}_2$ pour appliquer le théorème 2.1.4.

Le Tableau 2.1 représente le pourcentage de cas pour lesquels H_0 est rejetée. On constate que le choix de la partition est crucial pour les séquences les plus courtes. De plus, le taux de rejet dépend de l'alternative de H_0 . En effet, la statistique de Pearson donne les meilleurs résultats lorsqu'il s'agit de distinguer une petite séquence i.i.d. d'une séquence markovienne d'ordre 1. Pour des séquences plus longues, les taux de rejet sont équivalents. Comme le test de Pearson, les tests basés sur la CGR semblent posséder de bonnes propriétés asymptotiques.

Par ailleurs, deux partitions de même taille (i.e. avec le même nombre de sous-ensembles) n'ont pas nécessairement le même taux de rejet empirique ; on peut comparer la partition $\{B_*, \bar{B}_*\}$ avec \mathcal{P}_1 , ou comparer \mathcal{P}_2 et Pearson, ou encore \mathcal{P}_3 et \mathcal{P}_4 . \mathcal{P}_3 est meilleure pour distinguer des séquences i.i.d. de chaînes de Markov d'ordre $m > 1$.

On constate également que le taux de rejet de chaînes de Markov de longue dépendance augmente lorsque le nombre d'ensembles constituant la partition augmente. Pour des chaînes de Markov d'ordre 5, le taux de rejet du test construit avec \mathcal{P}_2 est semblable à celui du test de Pearson, alors qu'avec les partitions de 9 zones, le taux de rejet est plus grand.

Les statistiques définies dans les théorèmes 2.1.4 et 2.1.10 sont plus générales que la statistique de Pearson (2.5) car elles ne se limitent pas à une structure markovienne particulière. Le test de Pearson n'est pas consistant contre toutes les alternatives possibles (stationnaires ergodiques) contrairement aux tests basés sur la CGR. Pour illustrer ce point, on peut considérer la famille suivante de chaînes de Markov *mixées* d'ordre $m > 1$: dans un premier temps, on génère m chaînes de Markov indépendantes $U^{(1)}, \dots, U^{(m)}$ d'ordre 1 comme ci-dessus ; puis la séquence finale U est obtenue par l'agrégation

$$u_{km+i} = u_k^{(i)}, \text{ pour tout } k \geq 0 \text{ et } 1 \leq i \leq m. \quad (2.17)$$

U est une chaîne de Markov d'ordre m , où chaque nucléotide u_i ne dépend seulement que du nucléotide u_{i-m} , et est indépendant des u_{i-k} pour $1 \leq k < m$. Cependant, du point de vue de la statistique du Chi-deux (2.5), U se comporte comme une séquence i.i.d. et la probabilité de rejet de H_0 pour ce processus particulier ne tend pas vers 1. Lorsque l'on considère ce cas particulier de chaînes de Markov d'ordre m , les résultats de simulations montrent que le test basé sur la CGR se comporte assez bien, alors que la statistique de Pearson donne comme prévu de très faibles taux de rejet. Cet exemple illustre la force de la CGR pour ces tests : ils n'imposent pas de contrainte sur la séquence d'entrée plus forte que la stationnarité. L'alternative de ce test est plus large que H_1 . Un test de Pearson de H_{m-1} contre H_m permettrait certes de rejeter ces chaînes de Markov *mixées* d'ordre m , mais à partir du moment où le test accepte l'hypothèse d'indépendance (et même toutes les hypothèses H_j pour $j \leq m - 1$), on ne cherche a priori pas à continuer tous les tests d'ordre supérieur jusqu'à mettre en évidence un ordre de dépendance. L'avantage incontestable des tests basés sur la CGR est qu'il n'y a pas de restriction dans l'alternative (autre que stationnaire ergodique).

En résumé, certaines partitions sont plus robustes (les plus grosses) mais leur niveau converge plus lentement vers le niveau asymptotique. Ce qui confirme les assertions énoncées dans la Remarque 2.1.7.

Pour contourner cette difficulté, on va considérer le test défini dans le théorème 2.1.10 et le comparer numériquement à celui du théorème 2.1.4 sur plusieurs types de séquences. Les résultats sont donnés dans le Tableau 2.2.

La puissance du test construit à partir d'une collection de partitions est comparable à la puissance du meilleur des tests construits sur les partitions elles-mêmes. En mélangeant plusieurs partitions, on peut même améliorer la performance du test (par exemple

ordre	n	Pearson	$\{B_*, \bar{B}_*\}$	\mathcal{P}_1	\mathcal{P}_2	\mathcal{P}_3	\mathcal{P}_4
0	100	4.2	1.4	3.6	3.4	3.2	2.5
	500	6.1	1.8	4.8	4.8	4.6	3.8
	1 000	5.0	2.1	5.6	6.2	5.9	5.3
	10 000	6.5	1.8	4.8	5.1	5.9	7.7
1	100	86.4	21.3	12.9	51.1	44.5	28.9
	500	100	75.1	54.2	98.7	97.9	94.5
	1 000	100	84.6	70.9	99.9	99.5	99.0
	10 000	100	99.2	97.6	100	100	100
5	1 000	8.6	2.7	6.8	8.6	8.2	8.4
	10 000	54.6	18.5	28.7	55.6	81.5	85.3
	80 000	99.4	80.5	84.5	99.6	100	100
2 mixées	500	5.8	18.4	16.5	49.9	56.5	76.8
	1 000	7.0	35.7	26.9	73.7	83.0	95.1
	10 000	7.3	90.7	73.2	99.8	99.8	100
5 mixées	80 000	5.8	46.2	29.7	76.7	81.0	85.8
7 mixées	1 000 000	5.9	35.9	28.5	67.3	72.4	90.6

TAB. 2.1: Taux de rejet (en %) de H_0 , pour un niveau asymptotique $\alpha = 0.05$, en appliquant le test défini dans le théorème 2.1.4. La ligne *ordre* 0 représente le taux de rejet de séquences i.i.d. Sur les lignes *ordre* m on lit les taux de rejet de chaînes de Markov d'ordre m ($m \geq 1$). Pour comparaison, on ajoute les résultats du test de Pearson.

ordre	n	$\{\mathcal{P}, \mathcal{P}_2\}$	$\{\mathcal{P}, \mathcal{P}_4\}$	$\{\mathcal{P}, \mathcal{P}_2, \mathcal{P}_4\}$
0	100	3.9 (<i>4.8/5.3</i>)	4.8 (<i>4.0/5.3</i>)	3.7 (<i>4.0/5.3</i>)
	500	4.7 (<i>4.8/5.0</i>)	5.5 (<i>4.8/5.0</i>)	4.9 (<i>4.8/5.0</i>)
	1 000	4.0 (<i>4.9/5.0</i>)	5.8 (<i>5.0/5.8</i>)	4.9 (<i>4.9/5.8</i>)
	10 000	4.3 (<i>4.7/5.0</i>)	4.7 (<i>4.7/5.0</i>)	4.4 (<i>4.7/5.0</i>)
1	100	83.0 (<i>54.8/86.4</i>)	82.3 (<i>39.2/86.4</i>)	81.3 (<i>39.2/86.4</i>)
	500	99.7 (<i>97.7/99.8</i>)	99.8 (<i>96.3/99.8</i>)	99.8 (<i>96.3/99.8</i>)
	1 000	100 (<i>99.9/100</i>)	100 (<i>99.6/100</i>)	100 (<i>99.6/100</i>)
	5 000	100 (<i>100/100</i>)	100 (<i>100/100</i>)	100 (<i>100/100</i>)
5	1 000	8.3 (<i>8.5/8.9</i>)	10.6 (<i>8.5/10.7</i>)	7.4 (<i>8.5/10.7</i>)
	10 000	61.7 (<i>54.5/55.0</i>)	84.1 (<i>55.0/83.5</i>)	79.6 (<i>54.5/83.5</i>)
	80 000	100 (<i>99.5/99.5</i>)	100 (<i>99.5/100</i>)	100 (<i>99.5/100</i>)
2 mixées	500	41.0 (<i>7.6/48.6</i>)	72.6 (<i>7.6/78.2</i>)	70.3 (<i>7.6/78.2</i>)
	1 000	66.4 (<i>5.9/73.2</i>)	92.1 (<i>5.9/94.3</i>)	91.1 (<i>5.9/94.3</i>)
	10 000	99.7 (<i>6.8/99.9</i>)	100 (<i>6.8/100</i>)	100 (<i>6.8/100</i>)

TAB. 2.2: Taux de rejet de H_0 (en %) pour un niveau asymptotique $\alpha = 0.05$, et pour plusieurs collections de partitions arbitraires. Pour chaque partition, on teste H_0 avec la statistique définie dans le théorème 2.1.4. Les α_j sont choisis selon la deuxième procédure (2.16). La quantité γ est estimée par simulations. La partition \mathcal{P} est définie par $\mathcal{P} \stackrel{\text{def}}{=} \{Su, u \in \mathcal{A}\}$. Les valeurs associées aux partitions donnant le meilleur et le pire taux de rejet sont ajoutées en italique.

pour les chaînes de Markov d'ordre 5 de longueur 80 000, le taux de rejet est 100% pour la première collection, alors que chaque test sur l'une de ces partitions donne un taux de rejet de 99,5%). Considérer une collection peut être une solution au problème crucial du choix d'une partition unique.

Remarque 2.1.12. La détermination de α_j nécessite du temps de calcul supplémentaire pour le test basé sur la deuxième procédure. Ce dernier est d'ailleurs plus puissant que le test basé sur la première procédure avec tous les α_j égaux à α/p . Un autre choix de α_j dans la première procédure reviendrait à mettre un poids sur chaque partition. En pratique, les résultats numériques obtenus avec la première procédure avec $\alpha_j = \alpha/p$ sont comparables aux résultats du Tableau 2.2. De plus, l'implémentation est plus simple et le temps de calcul un peu plus court.

Test pour les modèles markoviens

Le Tableau 2.3 représente le pourcentage de cas où l'hypothèse H_1 est rejetée pour un niveau $\alpha = 0.05$ et pour les partitions \mathcal{P}_2 et \mathcal{P}_4 représentées sur la Figure 2.1. Les résultats du test de Pearson ont été ajoutés à la liste. La statistique de Pearson pour le test de H_1 contre H_2 est définie (voir par exemple Dacunha-Castelle et Dufflo [23]) de la façon suivante :

$$X^2 \stackrel{\text{def}}{=} \sum_{u,v,w \in \mathcal{A}} \frac{\left(N(uvw) - N(uv)N(vw)/N(v) \right)^2}{N(uv)N(vw)/N(v)}.$$

Sans surprise, la statistique de Pearson donne les meilleurs résultats pour distinguer les chaînes de Markov d'ordre 1 des chaînes d'ordre 2, de même que les chaînes de Markov mixées d'ordre 2 définies en (2.17). Cependant le test basé sur la CGR est très efficace dans le cas particulier de chaînes de Markov mixées d'ordre 3. De plus, pour les chaînes de Markov d'ordre 4, les deux partitions contenant 4 ensembles sont équivalentes, tandis que la partition formée de 9 ensembles est meilleure.

2.1.7 Application à des séquences biologiques

Dans cette section, on applique la famille de tests basés sur la CGR à des vraies séquences d'ADN. On choisit arbitrairement plusieurs génomes. La Figure 2.4 représente la liste des séquences testées. On considère à la fois des parties non codantes et des séquences complètes. Pour chaque séquence génomique, et pour plusieurs longueurs n , on applique les tests des théorèmes 2.1.4 et 2.1.8 construits à partir des partitions $\{\mathcal{P}_2, \mathcal{P}_4, \mathcal{P}\}$.

Dans un premier temps, on compare la structure des séquences non codantes avec des séquences complètes de *Mus musculus* et d'*Homo Sapiens*. La Figure 2.2 représente la

ordre	n	\mathcal{P}_2	\mathcal{P}_4	Pearson
1	500	5.4	3.6	5.2
	1 000	5.2	5.4	4.1
	10 000	6.4	6.0	5.3
2	500	98.4	91.6	100
	1 000	99.9	99.6	100
	5 000	100	100	100
4	500	20.2	30.3	22.8
	1 000	44.3	69.7	51.0
	5 000	99.6	100	99.9
3 mixées	500	29.6	48.5	5.4
	1 000	55.6	79.3	8.1
	5 000	96.4	99.8	8.3

TAB. 2.3: Taux de rejet de H_1 (en %), pour un niveau asymptotique $\alpha = 0.05$, et pour le test défini dans le théorème 2.1.8.

Notation	Séquence	Genbank
homsa	Homo Sapiens Chromosome 2	Complète : NT_005403
	Homo Sapiens Chromosome 2	Non codante : 2988_NT_005058
mmus	Mus Musculus Chromosome 8	Complète : NT_078586
	Mus Musculus Chromosome 2	Non codante : 1567_NT_039206
ratn	Rattus Norvegicus Chromosome 6	NC_005105
bacc	Bacillus Cereus ATCC 14579	NC_004722
mace	Methanosarcina Acetivorans C2A	NC_003552
mlot	Mesorhizobium Loti MAFF303099	NC_002678

TAB. 2.4: Liste des séquences utilisées dans les expériences.

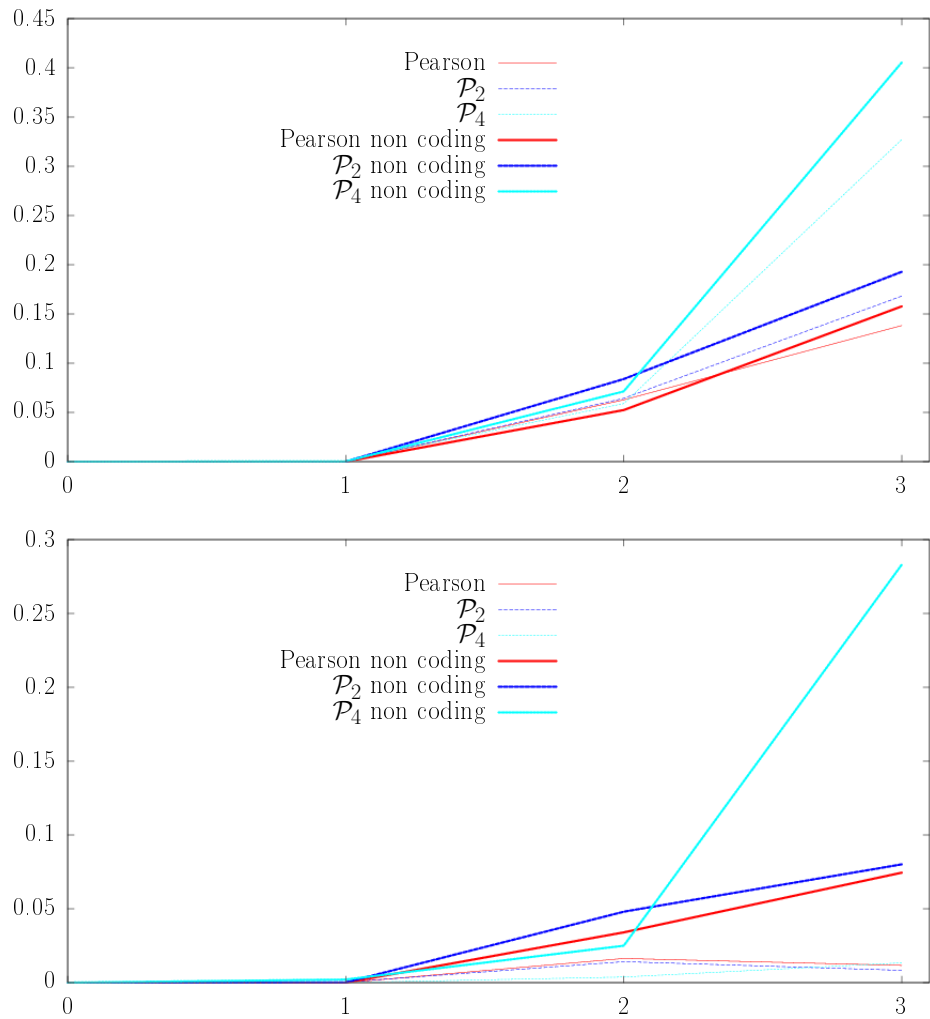


FIG. 2.2: Probabilité d'acceptation de H_m en fonction de m pour *Homo Sapiens* (en haut) et *Mus musculus* (en bas). On représente ici la moyenne prise sur 100 séquences de longueur 10 000.

probabilité d'acceptation

$$Q_j^m \left(\sum_{B \in \mathcal{P}, w, u} R_n^2(B, w, u) \right)$$

de H_m contre $H \setminus H_m$ en fonction de l'ordre m . Comme prévu, cette probabilité est plus grande pour des séquences non codantes ; cela confirme le caractère plus aléatoire des séquences non codantes, qui sont moins soumises à la pression sélective. De plus, pour les trois partitions \mathcal{P}_2 , \mathcal{P}_4 et \mathcal{P} , les probabilités d'acceptation sont du même ordre. Pour la statistique de Pearson, la différence entre les séquences complètes et les séquences non codantes est un peu moins prononcée. On remarque en particulier que pour les séquences d'*Homo Sapiens*, la statistique de Pearson donne la plus petite probabilité d'acceptation. D'un autre côté, pour *Mus musculus*, cette probabilité est minimale lorsque l'on choisit la partition \mathcal{P}_2 pour les séquences complètes. Pour l'ordre $m = 3$, la probabilité d'acceptation de \mathcal{P}_4 est un peu plus grande. En effet, par comparaison avec les partitions de 4 ensembles, il y a moins de points par ensemble dans \mathcal{P}_4 . Finalement, ces deux exemples confirment l'importance du choix de la partition. Une partition peut être bien adaptée à une alternative particulière mais donner d'un peu moins bons taux de rejet pour une autre alternative.

Dans une seconde série d'expériences sur les séquences d'*Homo Sapiens* avec les trois partitions \mathcal{P}_2 , \mathcal{P}_4 et \mathcal{P} , on calcule et représente dans la Figure 2.3 la probabilité d'acceptation en fonction de la longueur et de l'ordre m . Ici encore, les probabilités d'acceptation sont du même ordre de grandeur. Pour des séquences de longueur 50 000, avec le niveau $\alpha = 0.05$, H_m est rejetée pour $m \leq 4$ et acceptée pour $m \geq 5$ pour les trois partitions.

Pour le test de H_2 contre $H \setminus H_2$, quand la longueur est plus petite que 7 000, H_2 n'est pas encore rejetée quel que soit le choix de la partition. Pour des séquences de longueur plus grande que 6 000, la probabilité d'acceptation est minimale pour \mathcal{P}_4 .

Dans une dernière série d'expériences, on représente la probabilité d'acceptation pour plusieurs séquences de longueur 10 000 prises parmi plusieurs génomes, en fonction de m avec la partition \mathcal{P}_2 (voir le Tableau 2.4). Pour $m \leq 2$, H_m est rejetée pour toutes les espèces, dès que α est plus grand que 0.06. La probabilité d'acceptation est minimale pour *Bacillus Cereus* et maximale pour *Homo Sapiens*.

2.2 Preuves

2.2.1 Preuve du théorème 2.1.3

Lemme 2.2.1. *Sous l'hypothèse H_0 , le comportement asymptotique de la différence*

$$D_n(B, v) \stackrel{\text{def}}{=} \sqrt{n}(\hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B))$$

est donné par

$$D_n(B, v) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j(v) V_{j-1}(B) (1 + \eta_n),$$

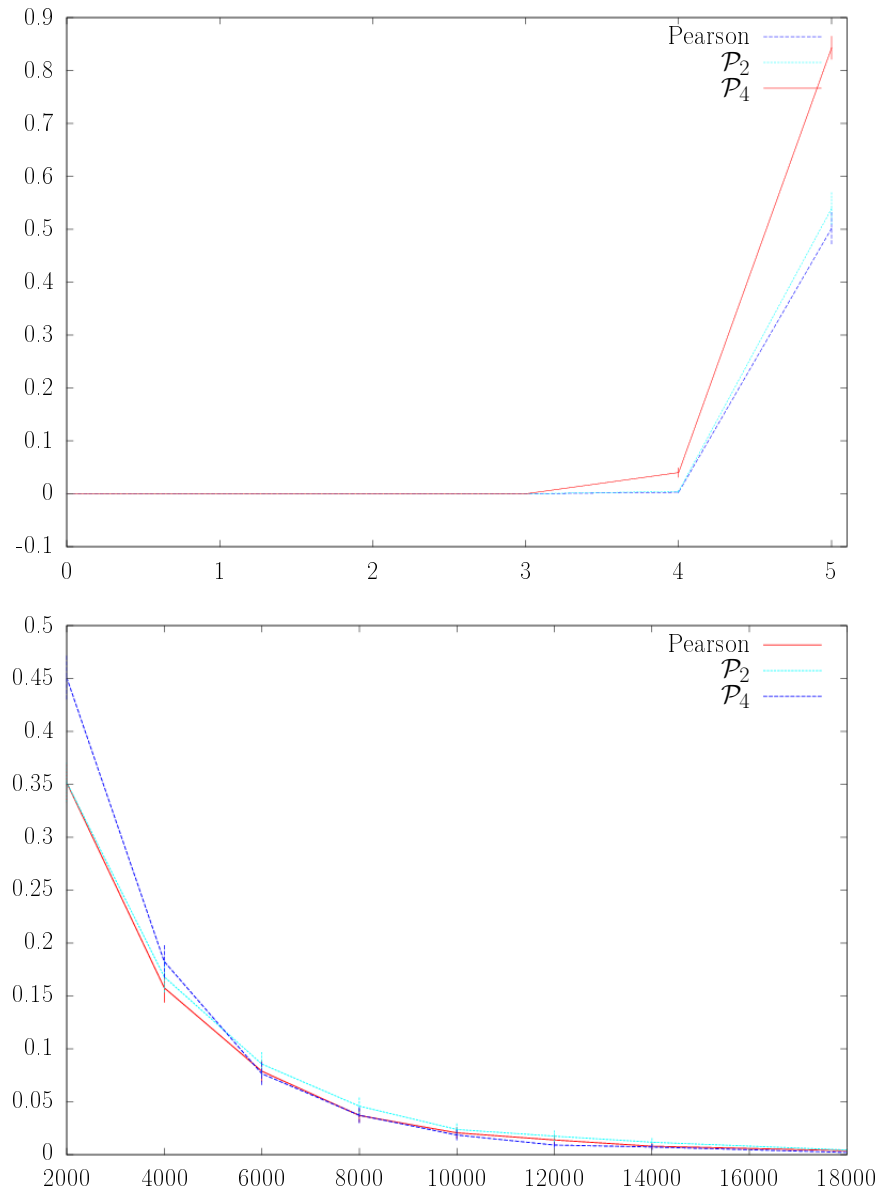


FIG. 2.3: Probabilité d'acceptation de H_m en fonction de m (en haut) et de la longueur n (en bas) pour des séquences d'*Homo Sapiens* et pour plusieurs partitions. On calcule la probabilité moyenne sur 1 000 séquences de longueur 50 000 et pour plusieurs ordres (en haut) (resp. pour l'ordre $m=2$ et différentes longueurs n (en bas)).

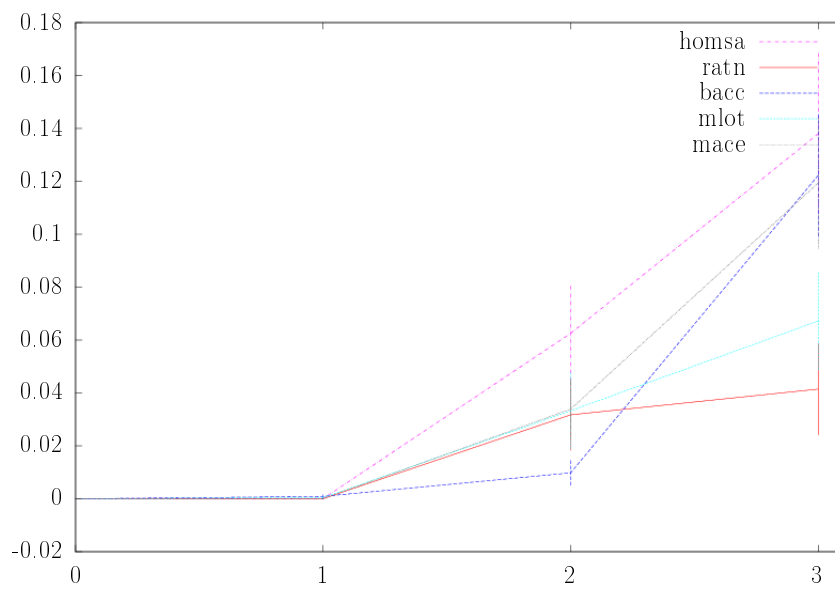


FIG. 2.4: Probabilité d'acceptation de H_m en fonction de m pour plusieurs espèces. On prend 200 séquences de longueur 10 000 dans la liste du Tableau 2.4, et on calcule la probabilité d'acceptation moyenne, pour chaque ordre, du test construit à partir de \mathcal{P}_2 .

avec

$$\lim_{n \rightarrow \infty} \eta_n = 0 \quad \text{en probabilité,}$$

où $\varepsilon_j(v) \stackrel{\text{def}}{=} \mathbb{1}_{\{u_j=v\}} - \mathbb{P}(u_j = v)$ et $V_j(B) \stackrel{\text{def}}{=} \mathbb{1}_{\{B\}}(X_j) - \pi(B)$.

Démonstration $D_n(B, v)$ peut se décomposer sous la forme

$$\begin{aligned} D_n(B, v) &= \sqrt{n} \left(\hat{\pi}_n(Bv) - (\hat{\pi}_n(Sv) - p_v)(\hat{\pi}_n(B) - \pi(B)) \right. \\ &\quad \left. - p_v \hat{\pi}_n(B) - \pi(B)(\hat{\pi}_n(Sv) - p_v) \right), \end{aligned}$$

où l'égalité $p_v \stackrel{\text{def}}{=} \mathbb{P}(u_1 = v) = \pi(Sv)$ est une conséquence du passage à la limite $n \rightarrow \infty$ dans l'égalité $\{X_{n+1} \in Sv\} = \{u_{n+1} = v\}$. De plus, sous H_0 , le théorème de la limite centrale classique établit la convergence

$$\sqrt{n}(\hat{\pi}_n(Sv) - p_v) \stackrel{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, p_v(1 - p_v)).$$

Ainsi, on déduit de la convergence de l'estimateur empirique et du lemme de Slutsky (voir par exemple Van der Vaart [88]) que

$$\lim_{n \rightarrow \infty} \sqrt{n} \left(\hat{\pi}_n(Sv) - p_v \right) \left(\hat{\pi}_n(B) - \pi(B) \right) = 0 \quad \text{en probabilité.} \quad (2.18)$$

Il reste à étudier le comportement asymptotique de la quantité

$$\begin{aligned} &\sqrt{n} \left[\hat{\pi}_n(Bv) - p_v \hat{\pi}_n(B) - \pi(B)(\hat{\pi}_n(Sv) - p_v) \right] \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \left[\mathbb{1}_{Bv}(X_j) - p_v \mathbb{1}_B(X_j) - \pi(B) \varepsilon_j(v) \right] \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \left[\mathbb{1}_{\{u_j=v\}} \mathbb{1}_B(X_{j-1}) - p_v \mathbb{1}_B(X_{j-1}) - \pi(B) \varepsilon_j(v) \right] \\ &\quad + \frac{1}{\sqrt{n}} p_v (\mathbb{1}_B(X_0) - \mathbb{1}_B(X_n)). \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j(v) V_{j-1}(B) + \frac{1}{\sqrt{n}} p_v (\mathbb{1}_B(X_0) - \mathbb{1}_B(X_n)). \end{aligned} \quad (2.19)$$

Puisque le deuxième terme de (2.19) tend vers 0 presque sûrement, le lemme 2.2.1 est une conséquence immédiate de (2.18) et (2.19). ■

Preuve du théorème 2.1.3. On déduit immédiatement du lemme 2.2.1, que

$$D_n(B, v) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j(v) V_{j-1}(B) (1 + \eta_n) \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} M_n(B, v) (1 + \eta_n), \quad (2.20)$$

où

$$\lim_{n \rightarrow \infty} \eta_n = 0 \quad \text{en probabilité.}$$

Soit $\mathbb{F} = (\mathcal{F}_n)$ la filtration naturelle du modèle avec $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$. La suite $M_n(B, v)$ est une \mathbb{F} -martingale de processus croissant

$$\langle M(B, v) \rangle_n = \sum_{j=1}^n \mathbb{E}[\varepsilon_j^2(v) \mid \mathcal{F}_{j-1}] V_{j-1}^2(B) = p_v(1-p_v) \sum_{j=1}^n V_{j-1}^2(B)$$

et, sous l'hypothèse H_0 , on voit aisément que

$$\sigma^2 \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} \frac{1}{n} \langle M(B, v) \rangle_n = p_v(1-p_v)\pi(B)(1-\pi(B)).$$

Puisque $M_n(B, v)$ a des accroissements bornés, la condition de Lindeberg est facilement vérifiée et le théorème de la limite centrale pour les martingales (voir théorème 4.1.4) entraîne que

$$\frac{1}{\sqrt{n}} M_n(B, v) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, \sigma^2). \quad (2.21)$$

Par conséquent, sous l'hypothèse H_0 , (2.21) et (2.20) impliquent

$$\sqrt{n} \left(\hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) \right) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, \sigma^2).$$

Puisque $\hat{\sigma}_n(u, B)$ converge presque sûrement vers σ , on déduit du lemme de Slutsky que

$$\frac{\sqrt{n}}{\hat{\sigma}_n(B, v)} \left(\hat{\pi}_n(Bv) - \hat{\pi}_n(Sv)\hat{\pi}_n(B) \right) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, 1)$$

et le théorème 2.1.3 est ainsi démontré. ■

2.2.2 Preuve du théorème 2.1.4

Lemme 2.2.2.

$$\sum_{B \in \mathcal{P}} \sum_{v \in \mathcal{A}} R_n^2(B, v) = \sum_{B \in \mathcal{P}} \sum_{v \in \mathcal{A}} \frac{1}{n} \left(\sum_{j=1}^n \frac{V_{j-1}(B)}{\sqrt{\pi(B)}} \frac{\varepsilon_j(v)}{\sqrt{p_v}} \right)^2 (1 + \eta_n),$$

où

$$\lim_{n \rightarrow \infty} \eta_n = 0 \quad \text{en probabilité.}$$

Démonstration On étudie le comportement asymptotique de

$$\begin{aligned} R_n(B, v) &= \frac{D_n(B, v)}{\sqrt{\hat{\pi}_n(Sv)\hat{\pi}_n(B)}} \\ &= \frac{D_n(B, v)}{\sqrt{p_v\pi(B)}} \frac{\sqrt{p_v\pi(B)}}{\sqrt{\hat{\pi}_n(Sv)\hat{\pi}_n(B)}}. \end{aligned}$$

Sous H_0 , les estimateurs empiriques $\hat{\pi}_n(B)$ et $\hat{\pi}_n(Sv)$ convergent presque sûrement vers $\pi(B)$ et p_v , respectivement, et donc

$$R_n(B, v) = \frac{D_n(B, v)}{\sqrt{p_v \pi(B)}} (1 + \zeta_n),$$

où

$$\lim_{n \rightarrow \infty} \zeta_n = 0 \quad \text{p.s.}$$

D'après le lemme 2.2.1,

$$\lim_{n \rightarrow \infty} R_n(B, v) - \frac{1}{\sqrt{n}} \sum_{j=1}^n \frac{V_{j-1}(B) \varepsilon_j(v)}{\sqrt{\pi(B)} \sqrt{p_v}} = 0 \quad \text{en probabilité,}$$

ce qui entraîne immédiatement le lemme 2.2.2. ■

Preuve du théorème 2.1.4 On définit tout d'abord les deux vecteurs colonnes

$$\begin{aligned} \xi_j &\stackrel{\text{def}}{=} \left(\frac{V_{j-1}(B)}{\sqrt{\pi(B)}} \right)_{B \in \mathcal{P}} \otimes \left(\frac{\varepsilon_j(v)}{\sqrt{p_v}} \right)_{v \in \mathcal{A}} \\ M_n &\stackrel{\text{def}}{=} \sum_{j=1}^n \xi_j, \end{aligned}$$

où $A \otimes B$ désigne le produit tensoriel entre A et B . La suite (M_n) est une \mathbb{F} -martingale de processus croissant

$$\langle M \rangle_n \stackrel{\text{def}}{=} \sum_{j=1}^n \mathbb{E} \left[\xi_j \xi_j^t \mid \mathcal{F}_{j-1} \right].$$

Sous l'hypothèse H_0 , il est facile de voir que pour tout $(u, v) \in \mathcal{A}$,

$$\mathbb{E} \left[\frac{\varepsilon_j(u)}{\sqrt{p_u}} \frac{\varepsilon_j(v)}{\sqrt{p_v}} \mid \mathcal{F}_{j-1} \right] = \mathbb{1}_{\{u=v\}} - \sqrt{p_u p_v}$$

et, pour tout $(B, B') \in \{B_1, \dots, B_K\}^2$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \frac{V_j(B)}{\sqrt{\pi(B)}} \frac{V_j(B')}{\sqrt{\pi(B')}} = \mathbb{1}_{\{B=B'\}} - \sqrt{\pi(B)\pi(B')}.$$

On définit ensuite les deux vecteurs colonnes

$$\begin{aligned} \sqrt{p} &\stackrel{\text{def}}{=} \left(\sqrt{p_u} \right)_{u \in \mathcal{A}} \\ \sqrt{\pi} &\stackrel{\text{def}}{=} \left(\sqrt{\pi(B)} \right)_{B \in \mathcal{P}}. \end{aligned}$$

On peut alors écrire

$$\lim_{n \rightarrow \infty} \frac{1}{n} \langle M \rangle_n \stackrel{\text{def}}{=} \Gamma = \left(I_K - \sqrt{\pi} \sqrt{\pi^t} \right) \otimes \left(I_d - \sqrt{p} \sqrt{p^t} \right) \quad \text{p.s.}$$

Γ est le produit tensoriel des matrices de projection orthogonale respectivement sur $(\sqrt{p})^\perp$ et sur $(\sqrt{\pi})^\perp$, en remarquant que $\sum_{B \in \mathcal{P}} \pi(B) = 1$.

Comme dans la preuve du théorème 2.1.3, la condition de Lindeberg est satisfaite et le théorème de la limite centrale pour les martingales entraîne

$$\frac{1}{\sqrt{n}} M_n \stackrel{\mathcal{L}}{\rightsquigarrow} \mathcal{N}_{dK}(0, \Gamma).$$

De plus, par continuité de la norme Euclidienne,

$$\frac{1}{n} \|M_n\|^2 \stackrel{\mathcal{L}}{\rightsquigarrow} \|Z\|^2, \quad \text{où } Z \stackrel{\mathcal{L}}{\rightsquigarrow} \mathcal{N}_{dK}(0, \Gamma),$$

ce qui implique

$$\sum_{B \in \mathcal{P}} \sum_{v \in \mathcal{A}} R_n^2(B, v) \stackrel{\mathcal{L}}{\rightsquigarrow} \|Z\|^2.$$

Finalement le théorème de Cochran nous permet d'établir $\|Z\|^2 \sim \chi^2(\delta)$ où le degré $\delta = \text{rang}(\Gamma)$. Dans le cas particulier où $K = 1$, on a $\Gamma = (1 - \pi(B))(I_d - \sqrt{p} \sqrt{p^t})$ et $\text{rang}(I_d - \sqrt{p} \sqrt{p^t}) = d - 1$, ce qui entraîne clairement (2.8). De plus, lorsque $K > 1$, $\delta = (d - 1)(K - 1)$ et ainsi on en déduit (2.10). ■

2.2.3 Preuve du théorème 2.1.8

Lemme 2.2.3. *Sous l'hypothèse H_m , le comportement asymptotique de la différence*

$$D_n(B, w, u) \stackrel{\text{def}}{=} \sqrt{n} \left(\hat{\pi}_n(Sw) \hat{\pi}_n(Bwu) - \hat{\pi}_n(Swu) \hat{\pi}_n(Bw) \right)$$

est donné par

$$D_n(B, w, u) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n-m-1} \varepsilon_{j+m+1}(w, u) V_j(B, w) (1 + \eta_n), \quad (2.22)$$

où

$$\lim_{n \rightarrow \infty} \eta_n = 0 \quad \text{en probabilité}$$

et

$$\begin{aligned} \varepsilon_j(w, u) &\stackrel{\text{def}}{=} \mathbb{1}_{Sw}(X_{j-1}) [\mathbb{1}_{\{u_j=u\}} - \mathbb{P}(u_{n+1} = u | X_n \in Sw)], \\ V_j(B, w) &\stackrel{\text{def}}{=} \pi(Sw) \mathbb{1}_B(X_j) - \pi(Bw). \end{aligned}$$

Démonstration La preuve est laissée au lecteur car elle est essentiellement semblable à celle du lemme 2.2.1. ■

Lemme 2.2.4.

$$\begin{aligned} & \sum_{wu \in \mathcal{A}^m \times \mathcal{A}, B \in \mathcal{P}} R_n^2(B, w, u) \\ &= \sum_{wu \in \mathcal{A}^m \times \mathcal{A}, B \in \mathcal{P}} \frac{1}{n} \left(\frac{1}{\sqrt{p_w}} \sum_{j=1}^{n-m-1} \frac{V_j(B, w)}{\sqrt{\pi(Bw)}} \frac{\varepsilon_{j+m+1}(w, u)}{\sqrt{p_{wu}}} \right)^2 (1 + \eta_n), \end{aligned}$$

où

$$\lim_{n \rightarrow \infty} \eta_n = 0 \quad \text{en probabilité.}$$

Démonstration La preuve est une conséquence directe du lemme 2.2.3, en utilisant la consistance forte de la mesure empirique. ■

Preuve du théorème 2.1.8. Les arguments sont semblables à ceux utilisés dans la preuve du théorème 2.1.4, en utilisant les vecteurs colonnes

$$\begin{aligned} \xi_{j+m+1} &\stackrel{\text{def}}{=} \left(\frac{1}{\sqrt{p_w}} \left(\frac{V_j(B, w)}{\sqrt{\pi(Bw)}} \right)_{B \in \mathcal{P}} \otimes \left(\frac{\varepsilon_{j+m+1}(w, u)}{\sqrt{p_{wu}}} \right)_{u \in \mathcal{A}} \right)_{w \in \mathcal{A}^m}, \\ M_n &\stackrel{\text{def}}{=} \sum_{j=1}^{n-d-1} \xi_j, \end{aligned}$$

où $p_w \stackrel{\text{def}}{=} \pi(Sw)$. On définit la matrice de transition Q pour tout $u \in \mathcal{A}$, et $w \in \mathcal{A}^m$ par

$$Q(w, u) \stackrel{\text{def}}{=} \mathbb{P}(u_{n+1} = u | X_n \in Sw).$$

La suite (M_n) est une \mathbb{F} -martingale de processus croissant

$$\langle M \rangle_n \stackrel{\text{def}}{=} \sum_{j=1}^n \mathbb{E} \left[\xi_j \xi_j^t \mid \mathcal{F}_{j-1} \right].$$

Alors, sous l'hypothèse H_m , on voit facilement que pour tous les $u, v \in \mathcal{A}$, $w \in \mathcal{A}^m$,

$$\mathbb{E} \left[\frac{\varepsilon_j(w, u)}{\sqrt{p_{wu}}} \frac{\varepsilon_j(w, v)}{\sqrt{p_{wv}}} \mid \mathcal{F}_{j-1} \right] = \frac{\mathbb{1}_{Sw}(X_{j-1})}{p_w} \left(\mathbb{1}_{\{v=u\}} - \sqrt{Q(w, u)Q(w, v)} \right)$$

et pour tout $(B, B') \in \mathcal{P}^2$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^{n-m-1} \frac{\mathbb{1}_{Sw}(X_{j+m})}{p_w^2} \frac{V_j(B, w)V_j(B', w)}{\sqrt{\pi(Bw)\pi(B'w)}} = \mathbb{1}_{\{B=B'\}} - \frac{\sqrt{\pi(Bw)\pi(B'w)}}{p_w}.$$

On définit les deux vecteurs colonnes

$$\begin{aligned}\sqrt{Q_w} &\stackrel{\text{def}}{=} \left(\sqrt{Q(w, u)} \right)_{u \in \mathcal{A}}^t \\ \sqrt{\pi_w} &\stackrel{\text{def}}{=} \left(\sqrt{\frac{\pi(Bw)}{p_w}} \right)_{B \in \mathcal{P}}^t.\end{aligned}$$

On montre comme précédemment que

$$\lim_{n \rightarrow \infty} \frac{1}{n} \langle M \rangle_n \stackrel{\text{def}}{=} \Gamma \quad \text{p.s.},$$

où Γ est une matrice carrée d'ordre $d^{m+1}K$, symétrique, définie positive diagonalisable par blocs avec des blocs de la forme $\Delta_w \otimes \Sigma_w$, où

$$\begin{aligned}\Sigma_w &\stackrel{\text{def}}{=} I_d - \sqrt{Q_w} \sqrt{Q_w}^{-t}, \\ \Delta_w &\stackrel{\text{def}}{=} I_K - \sqrt{\pi_w} \sqrt{\pi_w}^{-t}.\end{aligned}$$

Les matrices Σ_w et Δ_w sont les projections orthogonales respectivement sur $(\sqrt{Q_w})^\perp$ et sur $(\sqrt{\pi_w})^\perp$. La condition de Lindeberg est satisfaite et le théorème de la limite centrale pour les martingales s'applique et entraîne

$$\frac{1}{\sqrt{n}} M_n \stackrel{\mathcal{L}}{\rightsquigarrow} \mathcal{N}_{d^{m+1}K}(0, \Gamma).$$

Par continuité de la norme Euclidienne et par le théorème de Cochran, on en déduit que

$$\sum_{B \in \mathcal{P}} \sum_{wu \in \mathcal{A}^m \times \mathcal{A}} R_n^2(B, w, u) \stackrel{\mathcal{L}}{\rightsquigarrow} \|Z\|^2,$$

où $\|Z\|^2 \sim \chi^2(\delta)$ avec $\delta = d^m(d-1)(K-1)$ et la première partie du théorème 2.1.8 est ainsi démontrée. La puissance asymptotique du test est immédiate à partir de la convergence presque sûre de l'estimateur empirique ainsi que de la convergence suivante, comme dans la preuve du théorème 2.1.6,

$$\lim_{n \rightarrow \infty} \sum_{B \in \mathcal{P}} \sum_{wu \in \mathcal{A}^m \times \mathcal{A}} R_n^2(B, w, u) = +\infty \quad \text{p.s.}$$

■

2.3 Signature génomique et arbres taxonomiques

Deschavanne et al. [24] utilisent la CGR pour caractériser et classier les espèces. Ils utilisent les fréquences d'apparition de tous les mots, qui forment alors une « signature

génomique ». En effet, une analyse de fréquences le long d'un gène permet de mettre en évidence des similarités et des différences entre les espèces. Dans leur étude, la CGR n'est qu'un outil permettant de représenter ces signatures sous forme d'images, dans lesquelles les zones les plus foncées correspondent aux mots les plus fréquents. De plus, ils affirment que cette spécificité de la signature génomique, qui permet de « caractériser le style d'écriture », est une conséquence de l'action de l'environnement d'une part, et des structures de contraintes d'autre part.

Karlin et Burge [50], Karlin et Mrázek [52] utilisent un *profil de fréquences relatives de dinucléotides* comme signature génomique. Avec les notations de la Proposition 2.1.1, le rapport d'abondance relative du dinucléotide uv peut s'écrire sous la forme

$$\rho_{uv} \stackrel{\text{def}}{=} \frac{\pi(Suv)}{\pi(Su)\pi(Sv)}. \quad (2.23)$$

En pratique, on estime ces rapports d'abondance sur une partie du génome. Karlin et Mrázek [53], Campbell et al. [13] étudient le comportement de ce profil de fréquences relatives de dinucléotides pour différentes espèces d'ADN, c'est-à-dire le comportement de l'ensemble des

$$\hat{\rho}_{uv} \stackrel{\text{def}}{=} \frac{\hat{\pi}_n(Suv)}{\hat{\pi}_n(Su)\hat{\pi}_n(Sv)},$$

pour tous les couples de nucléotides (u, v) . Les mesures empiriques sont calculées à partir de séquences auxquelles on concatène leur complément inversé, afin que le choix particulier de l'un des deux brins n'influe pas sur les fréquences obtenues. Par exemple, la séquence *ATGCGAG* devient *ATGCGAGCTCGCAT*. Les valeurs du profil de fréquences relatives de dinucléotides sont équivalentes aux *general designs* dérivés des analyses de fréquences des plus proches voisins biochimiques (Josse et al. [49], Russel et Subak-Sharpe [81]).

Il semble remarquable que le profil d'abondance relative de dinucléotides ait un comportement stable dans le sens où, lorsqu'on le calcule sur des fenêtres de taille 50kb (i.e. 50 000 nucléotides) sur un génome donné, le profil est quasi identique à celui que l'on calculerait sur tout le génome de l'organisme (Karlin et Mrázek [53], Karlin et al. [54]). La stabilité de ce profil peut résulter de contraintes d'énergie qui sélectionnent certains dinucléotides, des pressions de mutations qui dépendent du contexte, des mécanismes de réplication et de réparation (Karlin et Mrázek [53], Karlin et al. [54], Karlin et Burge [50], Karlin et al. [51]). Dans une étude plus récente, Jernigan et Baran [48] testent l'hypothèse selon laquelle la sur- ou sous-représentation de dinucléotides dans un génome donné est invariante. Ils montrent que chaque terme du profil calculé sur une séquence de longueur n d'un génome donné, converge vers le profil calculé sur le génome entier à la vitesse $\log n$.

La comparaison de ces signatures génomiques permet de construire des arbres phylogénétiques. Il existe deux grands types de méthodes permettant la reconstruction d'arbres phylogénétiques : les méthodes basées sur les mesures de distances entre séquences et les

méthodes basées sur les caractères, qui prennent en compte les mutations au cours de l'évolution. Campbell et al. [13] étudient les avantages de l'utilisation de cette signature génomique. Les méthodes classiques sont basées sur des similarités et différences dans l'alignement de régions ou de gènes homologues. Les alignements de séquences relativement longues sont généralement difficiles à effectuer en pratique et on peut trouver des arbres phylogénétiques différents selon les séquences choisies pour la reconstruction. Grâce à leur stabilité locale, un arbre construit à partir de ces profils est indépendant de la région de longueur 50kb choisie dans la séquence. De plus, cette signature permet de prendre en compte à la fois des séquences codantes et des séquences non codantes.

Deschavanne et al. [24] montrent que la signature génomique peut, de manière plus précise, tirer avantage des fréquences des oligonucléotides de plus grandes tailles, et permettre de comparer des séquences homologues ou non. Par exemple, l'ensemble d'outils « GENSTYLE » (Fertil et al. [36]) permet de faire le rapprochement entre la signature du SRAS et celle des coronavirus.

Cette utilisation de la CGR conduit à se demander comment prendre en compte davantage d'information que le comptage de mots. L'idée de partitionner à nouveau l'ensemble sur lequel on construit la représentation reste valable dans ce contexte. En effet, à partir de la Proposition 2.1.1 et de la définition du profil (2.23), il est alors tentant de définir un *rapport d'abondance relative basé sur la CGR* par

$$\rho(B, v) \stackrel{\text{def}}{=} \frac{\pi(Bv)}{\pi(B)\pi(Sv)}, \quad (2.24)$$

qui vérifie trivialement $\rho(Su, v) = \rho_{uv}$.

Cette section a pour objectif de tester la performance de ce nouveau rapport. Pour cela, on choisit plusieurs séquences répertoriées dans le Tableau 2.5 (voir Figure 2.5 pour l'arbre taxonomique correspondant). On extrait des sous-séquences de longueur 100kb et on leur concatène leur complément inversé. Le profil d'abondance relative basé sur la CGR (2.24) est estimé par le profil empirique

$$\hat{\rho}(B, v) \stackrel{\text{def}}{=} \frac{\hat{\pi}_n(Bv)}{\hat{\pi}_n(B)\hat{\pi}_n(Sv)}.$$

2.3.1 Matrices de distances entre espèces

Pour une espèce Σ (resp. Σ') dont on considère n (resp. n') séquences numérotées arbitrairement, on note $\hat{\rho}_i(B, v)$ (resp. $\hat{\rho}'_i(B, v)$) le rapport d'abondance relative basé sur la CGR de la $i^{\text{ème}}$ séquence. Pour une partition donnée \mathcal{P} du carré unité, la *différence d'abondance relative basée sur la CGR* est définie pour deux espèces $\Sigma \neq \Sigma'$ par

$$\delta(\Sigma, \Sigma') = \frac{1}{4|\mathcal{P}||\Sigma||\Sigma'|} \sum_{i,j, B \in \mathcal{P}, v \in \mathcal{A}} |\hat{\rho}_i(B, v) - \hat{\rho}'_j(B, v)|$$

Abbr	Séquence	GenBank
homsa1	Homo Sapiens	NT_022184.13
homsa2	Homo Sapiens	NT_005403.14
homsa3	Homo Sapiens	NT_025741.13
homsa4	Homo Sapiens	NT_011520.9
homsa5	Homo Sapiens	NT_011757.13
mmus	Musmusculus	NT_07586.1
ratn1	Rattus Norvegicus	NC_005118
ratn2	Rattus Norvegicus	NC_005117
ratn3	Rattus Norvegicus	NC_005107
ratn4	Rattus Norvegicus	NC_005105
gall1	Gallus gallus	NC_006097.1
gall2	Gallus gallus	NC_006096.1
gall3	Gallus gallus	NC_006095.1
gall4	Gallus Gallus	NC_006094.1
gall5	Gallus Gallus	NC_006093.1
gall6	Gallus Gallus	NC_006092.1
gall7	Gallus Gallus	NC_006091.1
agam1	Anopheles gambiae	NW_045719.1
agam2	Anopheles gambiae	NW_045746.1
agam3	Anopheles gambiae	NW_045763.1
agam4	Anopheles gambiae	NW_045815.1
dme1a1	Drosophila melanogaster	NC_004354.1
dme1a2	Drosophila melanogaster	NT_033779.2
dme1a3	Drosophila melanogaster	NT_033778.1
dme1a4	Drosophila melanogaster	NT_037436.1
dme1a5	Drosophila melanogaster	Arm X
dme1a6	Drosophila melanogaster	Arm2R
dme1a7	Drosophila melanogaster	Arm 2L
dme1a8	Drosophila melanogaster	Arm3L
dme1a9	Drosophila melanogaster	Arm4
dme1a10	Drosophila melanogaster	Arm3R

Abbr	Séquence	GenBank
celeg1	Caenorhabditis elegans	CHR_I
celeg2	Caenorhabditis elegans	CHR_II
celeg3	Caenorhabditis elegans	CHR_III
celeg4	Caenorhabditis elegans	CHR_IV
celeg5	Caenorhabditis elegans	CHR_V
celeg6	Caenorhabditis elegans	CHR_X
plal	Plasmodium falciparum	NC_004317
y1lp1	Yarrowia Lipolytica	NC_006072
y1lp2	Yarrowia Lipolytica	NC_006071
y1lp3	Yarrowia Lipolytica	NC_006070
y1lp4	Yarrowia Lipolytica	NC_006069
osat1	Oryza Sativa	NT_036323
osat2	Oryza Sativa	NT_079973
osat3	Oryza Sativa	NT_080060
osat4	Oryza Sativa	NT_080067
osat5	Oryza Sativa	NT_080068
athal1	Arabidopsis thaliana	NC_003070
athal2	Arabidopsis thaliana	NC_003071.3
athal3	Arabidopsis thaliana	NC_003074.4
athal4	Arabidopsis thaliana	NC_003075.3
athal5	Arabidopsis thaliana	NC_003076.4
bacc	Bacillus cereus	NC_004722
braj	Bradyrhizobium japonicum	NC_004463
ccre	Caulobacter crescentus	NC_002636
mlot	Mesorhizobium loti	NC_002678
mbor	Mycobacterium bovis	AF2122/97
save	Streptomyces Avermitilis	NC_003155
scoe	Streptomyces Coelicolor	NC_003888
mace	Methanosarcina Acetivorans C2A	NC_003552
maze	Methanosarcina Mazei	NC_003901
ssol	Sulfolobus Solfataricus P2	NC_002754.1

TABLE 2.5: Liste des séquences utilisées dans les simulations pour la signature génomique. Toutes ces séquences sont disponibles sur le site <http://www-rocq.inria.fr/~cenac/sequences.html>

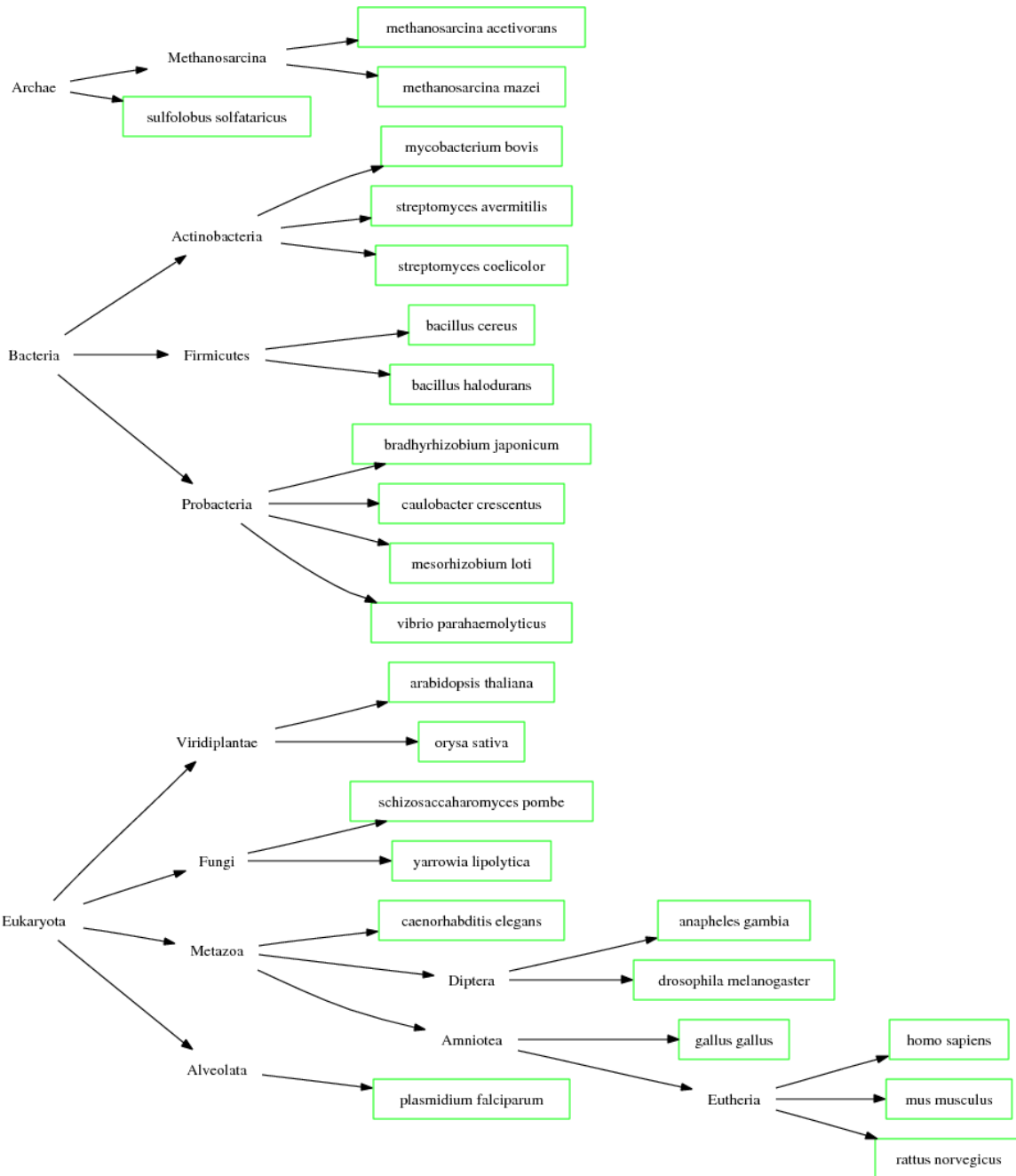


FIG. 2.5: Arbre taxonomique communément admis des espèces utilisées dans les simulations.

et dans le cas particulier où $\Sigma = \Sigma'$

$$\delta(\Sigma, \Sigma) = \frac{1}{4|\mathcal{P}||\Sigma|(|\Sigma| - 1)} \sum_{i,j,B \in \mathcal{P}, v \in \mathcal{A}} |\hat{\rho}_i(B_k, v) - \hat{\rho}_j(B_k, v)|.$$

Quand la partition utilisée est $\{Su, u \in \mathcal{A}\}$, δ coïncide avec la *différence d'abondance relative de dinucléotides* définie par Karlin et Mrázek [53]. Elle est calculée dans le Tableau 2.7. On peut la comparer à la distance d'abondance relative basée sur la CGR (donnée dans le Tableau 2.6) et calculée à partir de la partition \mathcal{P}_2 représentée sur la Figure 2.1. Cette partition est elle aussi formée de 4 ensembles, mais aucun de ses ensembles n'est réunion de carrés Sw correspondant à des mots w de « longueurs entières ».

Dans les deux cas, les différences entre séquences d'une même espèce sont très faibles. De plus, les familles *Eutheria*, *Diptera*, *Plantes*, *Firmicutes*, *Pro bacteria*, *Actinobacteria* et *Archae* forment des groupes cohérents. On compare tout d'abord les différences maximales intra-espèces avec les différences minimales entre espèces différentes. On peut remarquer que 14 espèces sont plus proches de leur propre groupe dans le cas des simulations basées sur la CGR, alors que seulement 10 espèces le sont pour les distances basées sur les rapports de fréquences de dinucléotides. Plus précisément, la démarcation est plus radicale pour les espèces suivantes : *Streptomyces Coelicolor*, *Streptomyces Avermitilis*, les *Bacillus*, *Plasmodium falciparum*, *Yarrowia Lipolytica*.

Afin de mesurer l'influence de la taille de la partition, on refait les mêmes calculs sur une partition \mathcal{P}' régulière de taille 10×10 zones, ce qui revient, selon la terminologie d'Almeida et al. [3] à compter les mots de longueur fractionnaire $\log_2 10 \approx 3.32$. Le Tableau 2.8 donne les différences moyennes δ (multipliées par 1 000) entre tous les rapports dans la partition \mathcal{P}' . Le nombre (17) d'espèces ayant une différence intra-groupe maximale supérieure à la différence minimale avec les autres groupes est plus grand que pour les autres partitions. En particulier, *Drosophila Melanogaster* et *Oryza Sativa* sont démarquées.

En conclusion, on constate, en comparant ces matrices de distances, que l'utilisation de zones qui ne correspondent pas à des mots, ni à des unions de mots, permet un gain de précision et d'information dans la comparaison des séquences biologiques. Les avantages évoqués par Fertil et al. [36] sont toujours valables dans l'étude de ces profils sur la CGR, à savoir la stabilité locale et le traitement rapide et simultané d'un grand nombre de séquences.

2.3.2 Arbres taxonomiques

La phylogénie cherche à étudier la formation et l'évolution d'organismes biologiques afin d'établir leur parenté. L'utilisation des signatures génomiques ne prend pas en compte à proprement parler les différents taux de mutation au cours de l'évolution. Les arbres obtenus à partir des matrices de mesures de différences par Karlin et Mrázek

	mace	maze	ssol	mbov	save	scoe	bacc	bach	braj	ccre	mlot	pfal	ylip	celeg	gal	homsa	mmus	ratn	agam	dmela	athal	osat
	(40)	(40)	(29)	(40)	(41)	(41)	(40)	(40)	(41)	(40)	(41)	(32)	(43)	(22)	(68)	(53)	(43)	(77)	(10)	(45)	(28)	(42)
échantillons																						
mace	23	28	103	162	173	173	80	68	139	111	130	140	115	91	135	114	144	140	116	106	94	87
maze		18	95	177	187	187	84	82	151	119	140	140	124	103	126	110	138	136	126	111	102	88
ssol	103		24	189	185	190	98	128	193	143	203	140	127	163	122	109	112	114	164	153	114	111
mbov				19	109	118	121	101	107	88	106	182	113	126	198	200	202	193	107	143	133	148
save				109	36	37	151	137	130	112	150	240	135	184	224	230	223	215	182	212	161	199
scoe				118	37	29	158	144	144	122	165	246	140	187	228	235	227	219	186	220	165	208
bacc							27	49	130	104	129	101	89	88	118	106	133	129	83	78	75	61
bach							17	17	104	85	101	129	86	69	135	123	150	146	77	83	76	76
braj									31	70	43	213	155	148	204	201	223	219	168	171	161	157
ccre										20	78	189	116	134	156	156	168	165	142	152	127	139
mlot											22	203	151	135	197	191	219	215	156	155	154	149
pfal												36	142	131	156	127	164	160	131	106	122	93
ylip													28	110	104	110	105	97	106	122	52	119
celeg														54	151	137	164	162	74	80	98	90
gal															34	60	47	50	147	145	87	135
homsa																37	61	60	137	125	88	104
mmus																	27	34	162	166	94	148
ratn																		34	156	163	87	146
agam																			43	67	98	90
dmela																			<i>67</i>	36	108	<i>57</i>
athal													<i>52</i>								34	<i>96</i>
osat																				<i>57</i>		28

TAB. 2.6: Différences δ d'abondance relative basée sur la CGR (multipliée par 1 000) entre les espèces représentées dans le Tableau 2.5, construites en utilisant la partition \mathcal{P}_2 . Lorsque la différence maximale intra-groupe est bien plus petite que toutes les différences entre groupes, les valeurs correspondantes sont en gras. Sinon, les valeurs problématiques sont en italique.

	mace	maze	sol	mbov	save	score	bacc	bach	braj	cre	mlot	pfal	yhip	celeg	gal	homsa	mmus	ratn	agam	dmela	athal	osat
échantillons	(40)	(40)	(29)	(40)	(41)	(41)	(40)	(40)	(41)	(40)	(41)	(32)	(43)	(22)	(68)	(53)	(43)	(77)	(10)	(45)	(28)	(42)
mace	26	32	<i>108</i>	202	208	209	109	89	203	164	184	168	122	102	166	131	156	149	148	127	83	90
maze		20	<i>102</i>	209	224	225	113	97	213	174	193	172	129	113	152	119	148	142	153	132	87	95
sol	108	102	31	243	246	248	145	171	267	218	272	147	151	183	165	124	136	131	212	188	125	123
mbov				20	118	<i>134</i>	143	122	130	<i>97</i>	115	195	132	147	237	246	253	241	114	162	157	155
save				<i>134</i>	36	36	175	150	154	<i>112</i>	168	254	143	183	277	280	279	268	172	217	175	191
score				<i>134</i>		25	181	157	166	<i>122</i>	183	268	147	186	281	285	282	271	177	226	178	201
bacc							32	<i>71</i>	171	131	157	155	121	94	146	152	181	168	77	69	88	<i>48</i>
bach							<i>71</i>	20	135	99	114	184	121	<i>66</i>	179	183	206	198	78	87	97	84
braj									35	63	52	270	213	179	279	290	309	302	180	195	204	186
cre									19	69	23	236	162	144	233	242	257	250	148	170	153	140
mlot											23	256	207	151	266	278	302	295	150	167	190	175
pfal												38	174	183	224	178	199	193	193	168	161	128
yhip													27	129	148	149	144	131	125	141	64	123
celeg														59	172	173	191	183	90	95	102	102
gal															38	68	68	67	174	155	125	140
homsa																35	57	54	189	163	123	129
mmus																	31	39	215	196	132	160
ratn																		38	201	186	117	150
agam																			48	76	116	103
dmela							<i>69</i>												<i>76</i>	43	114	76
athal													<i>64</i>								34	86
osat							<i>48</i>														86	31

TABLE 2.7: Différences d'abondance relative de dinucléotides (multipliées par 1 000) entre les espèces décrites dans le Tableau 2.5. Lorsque la différence maximale intra-groupe est bien plus petite que toutes les différences entre groupes, les valeurs correspondantes sont en gras. Sinon, les valeurs problématiques sont en italique.

	mace	maze	ssol	mbov	save	scoe	bacc	bach	braj	ccre	mlot	pfal	ylip	celeg	gal	homsa	mmus	ratn	agam	dmela	athal	osat
	(5)	(5)	(3)	(5)	(6)	(6)	(5)	(5)	(6)	(5)	(6)	(4)	(8)	(22)	(68)	(53)	(8)	(77)	(10)	(45)	(28)	(7)
échantillons																						
mace	21	40	<i>171</i>	259	275	295	<i>156</i>	<i>156</i>	259	260	246	243	182	161	194	192	195	191	192	171	157	178
maze		18	<i>178</i>	280	294	313	<i>170</i>	174	274	275	262	251	197	176	185	188	191	189	206	182	170	187
ssol			42	284	306	325	<i>156</i>	193	304	302	309	240	190	212	205	204	202	198	224	216	158	186
mbov				19	<i>165</i>	191	205	186	<i>146</i>	167	148	307	191	209	312	325	324	317	194	213	211	233
save				165	56	65	249	223	198	183	215	332	214	253	334	352	342	335	252	280	245	279
scoe				191	65	42	273	248	221	202	239	352	237	275	354	372	363	356	275	304	267	297
bacc							28	106	229	239	227	226	157	146	194	220	224	217	129	133	138	150
bach								19	187	203	184	251	158	145	222	244	240	233	143	155	137	174
braj									33	120	70	347	246	235	335	353	351	346	235	248	245	261
ccre									23	120	28	345	234	245	330	342	332	328	256	261	243	269
mlot											28	342	244	226	328	343	341	336	232	235	245	256
pfal												45	221	237	262	242	240	235	243	235	215	213
ylip													28	148	191	197	183	176	157	154	99	146
celeg														65	195	200	206	200	123	115	131	134
gal															43	100	95	97	203	183	179	185
homsa																49	81	86	224	192	177	173
mmus																	38	46	222	201	168	185
ratn																		43	215	198	162	182
agam																			50	101	148	142
dmela																				35	146	125
athal													<i>99</i>								<i>37</i>	<i>120</i>
osat																					120	31

TAB. 2.8: Différences δ d'abondance relative basées sur la CGR (multipliées par 1 000) entre les espèces listées dans le Tableau 2.5, construites à partir d'une grille régulière de taille 10×10 formant une partition. Lorsque la différence maximale intra-groupe est bien plus petite que toutes les différences entre groupes, les valeurs correspondantes sont en gras. Sinon, les valeurs problématiques sont en italique.

[53] sont en fait plutôt des arbres taxonomiques, la taxonomie étant la science qui étudie la classification des êtres vivants. Les arbres de la Figure 2.6 sont construits à partir des matrices des Tableaux 2.7 et 2.6 respectivement, générés avec la méthode dite *Neighbor-Joining* (grâce à l'outil NJPLOT) introduite par Saitou et Nei [82], et développée par Perrière et Gouy [72]. Cette méthode consiste à regrouper (*clustering*) les éléments deux par deux en prenant d'abord les deux plus proches (notés A et B) dans la matrice de distances. Un nouvel élément, noté C , les remplace dans la matrice. Pour calculer la distance Δ entre C et un autre élément D , on calcule la moyenne

$$\Delta(C, D) = \frac{1}{2}(\Delta(A, D) + \Delta(B, D)).$$

On itère ainsi jusqu'à regrouper tous les termes. La Figure 2.7 représente l'arbre construit avec le Tableau 2.8.

Sur la Figure 2.8, on construit l'arbre taxonomique de toutes les séquences de *Bacteria* répertoriées dans le Tableau 2.5. Dans les deux expériences, les séquences sont regroupées par espèces sauf pour une séquence de *Streptomyces Coelicolor* et une de *Streptomyces Avermitilis*. Pour l'arbre construit à partir des différences δ basées sur la CGR, les trois groupes *Firmicutes*, *Pro bacteria* et *Actinobacteria* sont clairement séparés en trois familles distinctes.

Dans une dernière série d'expériences, on s'intéresse aux reconstructions basées sur des partitions un peu plus originales que des ensembles de rectangles. On partitionne le carré unité en une partition régulière de taille 20×20 zones. Puis on regroupe les zones aléatoirement en 16 ensembles. Chaque zone a une probabilité $1/16$ d'appartenir à chacun des ensembles. La Figure 2.9 représente la forme de l'arbre obtenu. Pour comparaison, l'arbre construit à partir de la partition régulière en 16 zones, équivalente au comptage de mots de 3 lettres, a été ajouté à la Figure 2.10. Les séquences utilisées pour ces expériences sont celles du Tableau 2.5 auxquelles on a ajouté celles du Tableau 2.9.

Dans l'arbre construit à partir des 400 zones groupées en 16 ensembles, les trois familles d'archées, eucaryotes et bactéries sont bien séparées (à l'exception de 3 espèces d'archées). Au contraire, dans l'arbre construit à partir des zones correspondant aux trinuécléotides, les espèces sont davantage mélangées.

Une nouvelle fois, le résultat est plus satisfaisant avec le profil basé sur la CGR ne correspondant pas à du comptage de mots.

Les arbres présentés ici sont des exemples de reconstruction taxonomique à partir de la CGR, qui donnent des résultats plus satisfaisants que les arbres construits à partir des profils de fréquences de dinuécléotides. On pourrait aussi penser à toutes sortes de partitions (diagonales, sinusoïdales, fractales, ...) où les zones ne sont pas rectangulaires. Les résultats ne sont pas tous présentés dans cette thèse, mais peuvent être consultés en ligne sur <http://mycgr.inria.fr/>. Les programmes qui ont permis ces expérimentations sont aussi sur ce site (cf. Section 2.4).

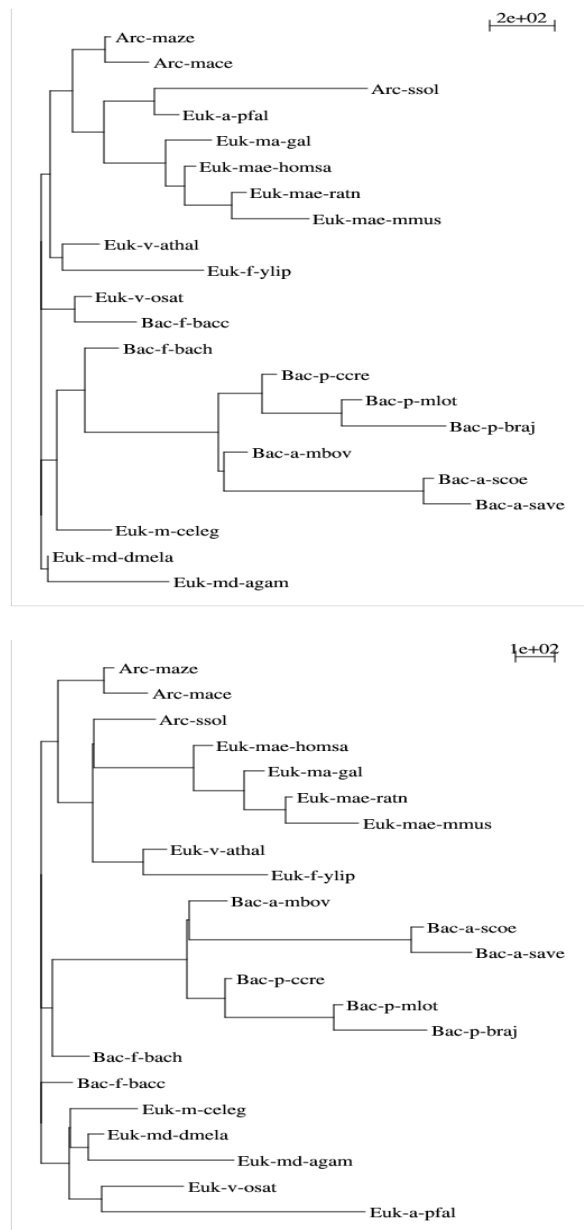


FIG. 2.6: Arbres taxonomiques non enracinés construits avec la méthode *Neighbor-Joining* à partir des différences d'abondance relative de dinucléotides du Tableau 2.7 (en haut) et des différences d'abondance relative basées sur la CGR du Tableau 2.6 (en bas). *Amniotea*, *Pro bacteria* et *Actinobacteria* forment des groupes cohérents. Au contraire, les groupes *Firmicutes*, *Metazoa*, *Viridiplantae* sont séparés et les *Archae* sont mélangées avec des *Eukaryota*.

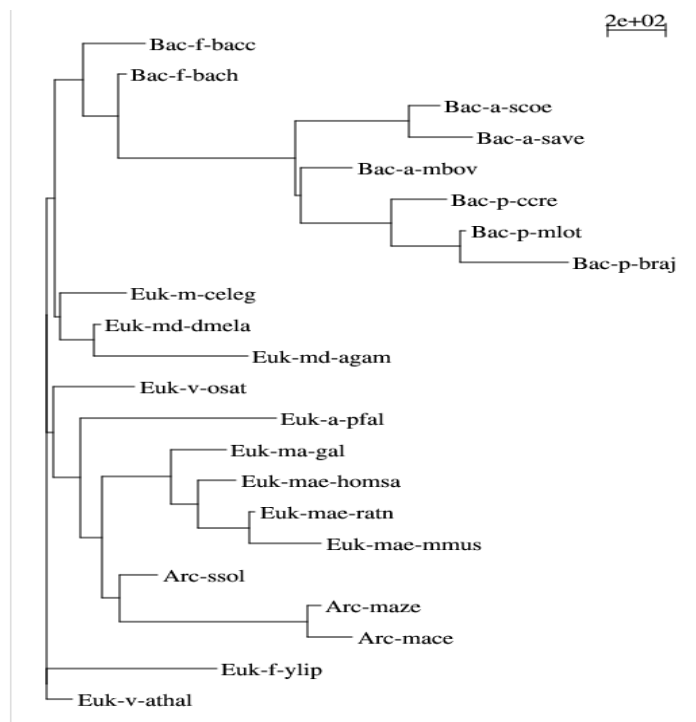


FIG. 2.7: Arbre taxonomique non enraciné construit avec la méthode *Neighbor-Joining* à partir des différences d'abondance relative basées sur la CGR du Tableau 2.8 avec la partitions régulière de taille 10×10 .

Abbr	Séquence	GenBank
paer	Pyrobaculum Aerophilum	NC_003364
stok	Sulfolobus Tokodaii	NC_003106
aful	Archaeoglobus Fulgidus	NC_000917
halo	Halobacterium sp	NC_002607
mkan	Met hanopyrus Kandleri	NC_003551
mt her	Met hanothermobacter Therautotrophicus	NC_000916
paby	Pyrococcus Abyssii	NC_000868
phor	Pyrococcus Horikoshii	NC_000961
taci	Thermoplasma Acidophilum	NC_002578
tvol	Thermoplasma volcanium	NC_002689
bt he	Bacteroides Thetaiotaomicron	NC_004663
viol	Chromobacterium Violaceum	NC_005085
ecol	Escherichia Coli	NC_004431
rbal	Rhodopirellula Baltica	NC_005027
vibp	Vibrio Parahaemolyticus	NC_004603
xcam	Xanthomonas Campestri	NC_003902
ypse	Yersinia Pseudotuberculosis	NC_006155

TAB. 2.9: Liste des séquences supplémentaires utilisées pour les résultats des Figures 2.9 et 2.10.



FIG. 2.8: Arbres taxonomiques non enracinés construits avec la méthode *Neighbor-Joining* à partir des différences d'abondance relative de dinucléotides du Tableau 2.7 (en haut) et des différences d'abondance relative basées sur la CGR du Tableau 2.6 (en bas) pour toutes les séquences de bactéries du Tableau 2.5. Avec la CGR, les 3 groupes *Firmicutes*, *Pro bacteria* et *Actinobacteria* sont démarqués, alors que *Firmicutes* apparaît comme une famille d'*Actinobacteria* dans l'arbre du haut.

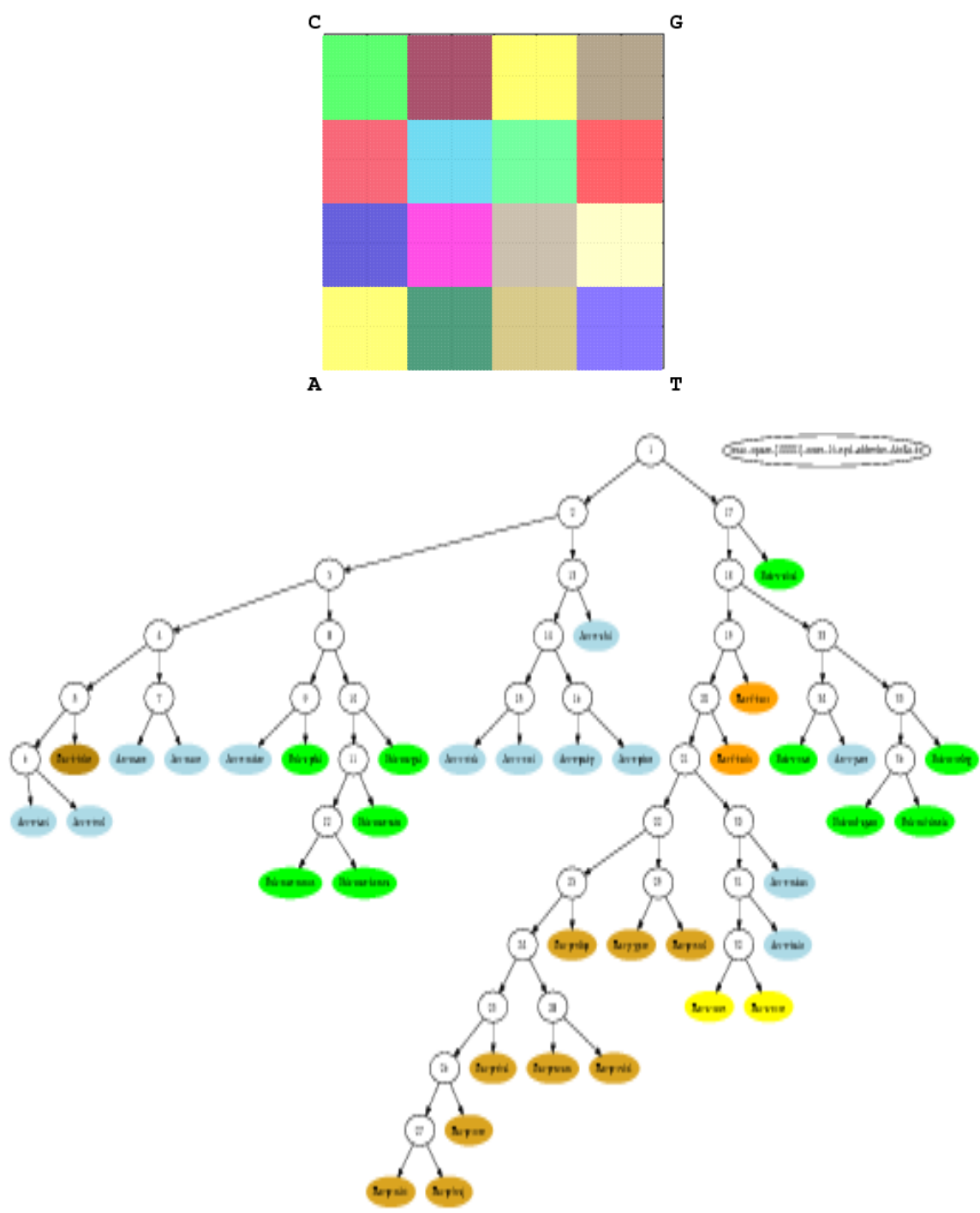


FIG. 2.10: Arbre taxonomique (en bas) construit à partir des différences d'abondance relative basées sur la CGR avec la partition régulière de 16 zones (en haut), correspondant au comptage des trinuécléotides.

2.4 Logiciels développés

Afin d'effectuer des simulations et mettre en pratique la nouvelle famille de tests ainsi que la signature génomique basée sur la CGR, une suite de programmes a été développée, en Objective-Caml (Leroy et al. [61]).

Elle se présente sous la forme d'une bibliothèque de modules et différents programmes. Le tout est mis à disposition comme logiciels libres sur le site <http://mycgr.inria.fr/>, avec une interface graphique permettant d'accéder facilement aux principales fonctionnalités. Ces développements ont été faits en collaboration avec Maxence Guesdon.

La bibliothèque de modules et les programmes permettent notamment :

- la manipulation de séquences : lecture/écriture dans des fichiers, découpage (pour effectuer des simulations sur des séquences de taille donnée), génération selon des lois i.i.d. ou markoviennes, inversion, . . . ,
- la construction et l'affichage de la CGR sur le segment, le carré, le tétraèdre : calcul des coordonnées à partir d'une séquence, gestion des zones et partitions (définition, affichage, lecture/écriture dans des fichiers, comptage des points),
- la manipulation de matrices, notamment les matrices de distances, et la création d'arbres par la méthode « Neighbor-Joining » à partir de ces matrices, avec génération de fichiers au format Graphviz (Ellson et al. [31]) et Newick à partir de ces arbres,
- la génération de tableaux L^AT_EX à partir de matrices,
- d'effectuer des simulations pour les tests définis dans la Section 2.1,
- d'effectuer les calculs de signature génomique ; les résultats sont présentés sous forme de matrices de distances et d'arbres aux formats Newick et Graphviz (cf. Section 2.3),
- le stockage de données dans une base pour ne pas les recalculer plusieurs fois et gagner du temps de calcul.

Nous avons choisi OCaml car le premier outil développé était utilisé pour représenter des séquences par des arbres. OCaml est particulièrement bien adapté pour manipuler des structures récursives, grâce au typage fort et au *pattern-matching*. De plus, il permet un développement très rapide de prototypes, ce qui convient à nos besoins de développer des programmes rapidement pour vérifier nos théories. Enfin, ce langage est développé à l'INRIA Rocquencourt, et la proximité de spécialistes du langage peut être d'une aide précieuse.

Présentons plus précisément en quelques lignes les fonctionnalités de chacun de ces programmes.

`mycgr_seq.x` permet de

- générer des séquences i.i.d. ou markoviennes, selon une loi et une longueur données en paramètre,
- calculer les fréquences empiriques des nucléotides dans une séquence,
- découper une séquence en plusieurs sous-séquences d'une longueur inférieure donnée.

`mycgr_square.x` effectue différents calculs sur les points (X_n) de la CGR dans le carré :

- calcul et affichage des coordonnées des points de la CGR, à partir d'une séquence donnée,
- comptage des points de la CGR dans différentes zones indiquées en paramètre, à partir d'une séquence donnée,
- calcul des différentes statistiques de tests définies dans le Chapitre 2.

Différentes options permettent de modifier le comportement du programme, par exemple en prenant des séquences déjà existantes pour faire les tests plutôt qu'en régénérer à chaque simulation. Cette option permet de comparer les différents tests en les appliquant aux mêmes séquences. On peut également indiquer des paramètres pour le *cache* afin d'optimiser les calculs dans certains cas. Les programmes `mycgr_segment.x` et `mycgr_tetra.x` offrent les mêmes fonctionnalités pour les points de la CGR sur le segment et dans le tétraèdre.

`mycgr_square_dist.x` implémente le calcul des distances entre séquences, par la généralisation dans la CGR du profil d'abondance relative de dinucléotides. Il existe plusieurs options pour le calcul de ces distances, en particulier on peut choisir de calculer les valeurs absolues ou les carrés des différences. De plus, on peut éventuellement regrouper les distances par espèces. Les résultats peuvent être générés aux formats Graphviz, Newick¹, PHYLIP², L^AT_EX. Les calculs sur plusieurs centaines de longues séquences et pour des partitions comportant de nombreuses zones prennent du temps. Pour pallier à ce problème, le programme peut les lancer en parallèle sur plusieurs machines en plaçant tous les fichiers sur un système de fichiers commun (NFS par exemple). Le cluster de l'INRIA Rocquencourt a ainsi été utilisé pour diviser par 15 les temps des longs calculs. Une base de données peut également être utilisée pour stocker plus efficacement les calculs intermédiaires. Les programmes `mycgr_segment_dist.x` et `mycgr_tetra_dist.x` offrent les mêmes services pour la CGR sur le segment et dans le tétraèdre.

`mycgr_square_draw.x` dessine la CGR dans le carré. Il lit les coordonnées des points à dessiner sur son entrée standard (ces coordonnées sont générées par le programme `mycgr_square.x`). Plusieurs options permettent :

- d'afficher une grille de taille donnée,
- de dessiner une partition décrite dans un fichier en paramètre,
- de montrer la construction des points avec des flèches,
- de cacher les coordonnées et/ou les lettres correspondants aux coins,

¹C'est notamment le format en entrée du programme NJPLOT utilisé pour dessiner les arbres comme ceux de la Figure 2.8. Le format est décrit à l'adresse suivante :

<http://evolution.genetics.washington.edu/phylip/newicktree.html>

²<http://evolution.genetics.washington.edu/phylip.html>

- d’afficher, plutôt que les points, les fréquences de points dans des sous-carrés de taille donnée. La couleur des sous-carrés est d’autant plus foncée que la fréquence des points y est élevée.

Les fichiers générés sont au format PostScript ou Embedded-PostScript.

`mycgr_square_test_markov.x` permet d’évaluer empiriquement le niveau et la puissance des tests de structure markovienne d’ordre m . En paramètre, on choisit le nombre d’expériences, la ou les partitions, les longueurs et les types de séquences (i.i.d., markoviennes, markoviennes mixées) que l’on veut tester. On peut également faire ces simulations avec la méthode de Bonferroni, selon les deux procédures décrites en Section 2.1.5. Les résultats sont générés au format L^AT_EX. Les programmes équivalents existent pour la CGR sur le segment (`mycgr_segment_test_markov.x`) et dans le tétraèdre (`mycgr_tetra_test_markov.x`).

`mycgr_square_vn.x` évalue empiriquement le niveau et la puissance du test d’indépendance, en choisissant le nombre d’expériences, les partitions, les longueurs et les types de séquences (i.i.d., markoviennes, markoviennes mixées) que l’on veut tester. Ce programme permet également de faire ces simulations avec la méthode de Bonferroni. Les résultats sont générés au format L^AT_EX. Les programmes équivalents existent pour la CGR sur le segment (`mycgr_segment_vn.x`) et dans le tétraèdre (`mycgr_tetra_vn.x`).

`mycgr_square_zones.x` génère des fichiers de zones ou partitions sur le carré unité. Il peut s’agir de zones correspondant à des mots, ou à des subdivisions régulières ou aléatoires du carré. On peut aussi générer des rectangles et des cercles aléatoires, sachant qu’il est toujours possible, pour faire une partition, d’indiquer l’une des zones comme étant le complémentaire des autres. Enfin, il est également possible de définir une partition en découpant le carré régulièrement ou aléatoirement en une multitude de zones et de les grouper en n ensembles, éventuellement de façon non équiprobable. Par exemple, la partition indiquée sur la Figure 2.9 a été créée en construisant 400 zones régulières sur le carré et en les groupant en 16 ensembles, chaque zone ayant la même probabilité 1/16 d’appartenir à l’un de ces ensembles. Les programmes équivalents existent pour la CGR sur le segment (`mycgr_segment_zones.x`) et dans le tétraèdre (`mycgr_tetra_zones.x`).

Du fait des nombreuses expériences menées sur différentes tailles de séquences avec différentes zones et différentes méthodes de calcul des distances, des conventions de nommage des fichiers sont devenues nécessaires. Tous les fichiers sont donc placés dans une arborescence de répertoires dont la racine (`meta_root`) est un paramètre de compilation. On organise alors les fichiers de la manière suivante :

`meta_root/sequences` contient les séquences originales.

`meta_root/size` contient les séquences de taille `size` isolées à partir des séquences originales.

`meta_root/cache` contient les fichiers de cache.

`meta_root/zones/segment` contient les fichiers de partitions sur le segment.

`meta_root/zones/square` contient les fichiers de partitions dans le carré.

`meta_root/zones/tetra` contient les fichiers de partitions dans le tétraèdre.

`meta_root/results/size/dists` contient les fichiers de résultats de calculs de distances avec des séquences de la taille *size*. Dans ce répertoire, les noms des fichiers suivent également une convention de nommage pour indiquer la méthode de calcul de distance utilisée, le fichier de partition utilisé, si le complément inversé a été ajouté aux séquences, et si la CGR était sur le segment, le carré ou le tétraèdre.

Le programme `mycgr_meta.x` permet de lancer les autres programmes de MyCGR en ajoutant les options et les noms de fichiers nécessaires pour respecter les conventions de nommage. Cela simplifie les commandes à lancer pour utiliser les programmes et ranger les résultats aux bons endroits.

`mycgr.x` offre une interface graphique donnant accès aux principales fonctionnalités des autres outils, en utilisant lui-aussi les conventions de nommage des fichiers :

- gestion des séquences originales et extraction de séquences plus petites pour les utiliser dans les simulations,
- affichage de la CGR sur le carré pour une séquence sélectionnée, ou un fichier de partition donné; on peut également superposer les deux représentations et ainsi visualiser la fréquence des points dans les zones d'une partition. L'utilisateur peut sauvegarder le résultat dans un fichier image,
- gestion des fichiers de zones définis pour le segment, le carré, le tétraèdre,
- parcours des résultats déjà obtenus et affichage des fichiers par l'outil correspondant (qui peut être paramétré),
- lancement du calcul des distances entre espèces, pour une longueur de séquence donnée et autres paramètres.

La Figure 2.11 montre une capture d'écran de `mycgr.x`.

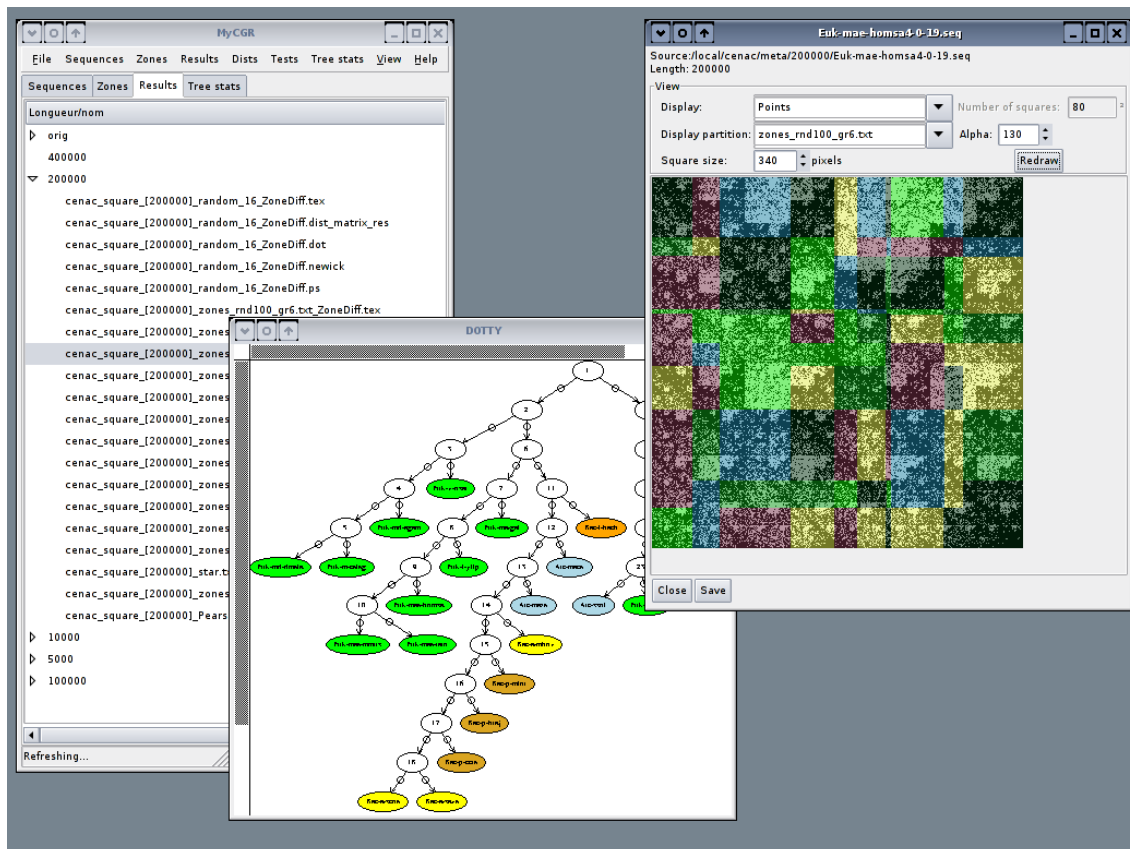


FIG. 2.11: Capture d'écran du programme `mycgr.x`. On distingue la fenêtre principale, à gauche, montrant les résultats disponibles. La fenêtre du milieu affiche un arbre taxonomique obtenu par la méthode de Neighbor-Joining et le calcul des différences d'abondance relative basées sur la CGR. La dernière fenêtre montre la CGR dans le carré pour une séquence d'*Homo Sapiens*, en superposition avec l'affichage d'une partition de 6 zones composées de 10×10 sous-carrés non réguliers répartis aléatoirement dans les zones.

Chapitre 3

Représentation d'une séquence d'ADN en arbre quaternaire

Le travail présenté dans ce chapitre est le fruit d'une collaboration avec B. Chauvin, S. Ginouillac et N. Pouyanne de l'université de Versailles-Saint-Quentin en Yvelines.

Nous proposons ici une représentation possible, inspirée par la CGR, de séquences d'ADN en arbres quaternaires. L'idée, à la base de la construction, est l'insertion successive dans un arbre digital de recherche de tous les *préfixes retournés* d'une séquence générée par une chaîne de Markov d'ordre un.

Nous démontrons que les longueurs des branches critiques et la profondeur d'insertion dans de tels arbres se comportent, au premier ordre, comme si les mots insérés formaient une famille de chaînes de Markov indépendantes les unes des autres.

Sommaire

3.1	Introduction	66
3.2	Construction de l'<i>arbre-CGR</i> et relation avec l'arbre digital de recherche	66
3.2.1	Algorithme de construction	66
3.2.2	Liens avec les arbres des suffixes	70
3.2.3	Occurrences de mots et recouvrements	73
3.2.4	État de l'art sur les arbres digitaux de recherche	74
3.3	Convergence presque sûre de branches critiques	75
3.3.1	Préambule	75
3.3.2	Lemme préliminaire	80
3.3.3	Minoration de la limite inférieure de la longueur des plus courtes branches	83
3.3.4	Majoration de la limite supérieure de la hauteur	85
3.4	Convergence en probabilité de la profondeur d'insertion	91
3.5	Expérimentations numériques	93
3.6	Logiciels développés	96
3.7	Domaine de définition de la fonction génératrice de la variable représentant la première occurrence d'un mot	96
3.7.1	Preuve de l'assertion ii) de la Proposition 3.3.3	96
3.7.2	Preuve de l'assertion i)	98

Les notations utilisées dans ce chapitre sont celles définies dans le Chapitre 1 de présentation de la CGR.

3.1 Introduction

Une propriété fondamentale de la CGR, abordée au Chapitre 1, est que tous les mots possédant un même suffixe $w = w_1 \dots w_d$ sont regroupés dans une même zone Sw . À partir d'une séquence U à valeurs dans un alphabet fini \mathcal{A} , on regroupe ainsi dans Sw tous les points X_n tels que $U_{n-d+1} \dots U_{n-1} U_n = w$. Comme chaque point de la CGR contient toute l'histoire de la séquence, on peut imaginer un instant que cette CGR n'est pas construite à partir d'une unique séquence U , mais plutôt de l'ensemble des mots

$$\begin{array}{c} U_1 \\ U_1 U_2 \\ \vdots \\ U_1 U_2 \dots U_n \\ \vdots \end{array}$$

L'appartenance de chacune de ces sous-séquences à une zone Sw donnée est déterminée par son suffixe. Une idée naturelle est de « ranger » ces sous-séquences dans les nœuds d'un arbre, tout en conservant la visualisation de répétitions de suffixes. Ci-après, nous proposons une représentation de séquences d'ADN dans des arbres quaternaires permettant cette conservation. Elle peut facilement être étendue à des séquences de lettres prises dans un alphabet fini quelconque.

3.2 Construction de l'arbre-CGR et relation avec l'arbre digital de recherche

3.2.1 Algorithme de construction

On supposera les lettres classées arbitrairement dans l'ordre (A, C, G, T) . On note \mathcal{T} l'arbre quaternaire infini complet. On peut imaginer cet arbre comme un canevas infini. À chaque étape de la construction, on insère un nœud dans ce canevas. On construit ainsi une suite de sous-arbres finis \mathcal{T}_n de \mathcal{T} , tous emboîtés $\mathcal{T}_1 \subset \mathcal{T}_2 \dots \mathcal{T}_n \subset \dots \subset \mathcal{T}$. Chaque sous-arbre \mathcal{T}_n possède n nœuds étiquetés (sans compter la racine), le processus d'étiquetage (expliqué ci-après) étant inhérent à la construction.

Étant donné une séquence aléatoire $U = U_1U_2\dots$, l'*arbre-CGR* pousse en insérant successivement des mots $U_1\dots U_i$ dans l'arbre infini complet. Chaque nœud de cet arbre possède 4 branches correspondant au quadruplet ordonné (A, C, G, T) .

Tout d'abord, la lettre U_1 est insérée dans l'arbre infini complet au niveau 1, juste sous la racine, sous la branche correspondant à U_1 . L'insertion du mot $U_1\dots U_n$ est faite récursivement de la manière suivante : on essaye tout d'abord de l'insérer au niveau 1 dans la branche correspondant à la dernière lettre rencontrée dans la lecture de la séquence, c'est-à-dire U_n ; si ce nœud est déjà occupé, on essaye de l'insérer au niveau 2 de l'arbre dans le sous-arbre correspondant à U_n , sous la branche correspondant à la lettre U_{n-1} . On répète l'opération jusqu'au premier nœud non occupé au niveau k dans la branche correspondant à la lettre U_{n-k+1} ; le mot $U_1\dots U_n$ est alors inséré sur ce nœud. Si $k < n$, seul compte le suffixe $U_{n-k+1}\dots U_n$ du mot $U_1\dots U_n$ que l'on insère.

Exemple de construction

La Figure 3.1 montre les premières étapes de la construction de l'arbre-CGR correspondant à toute séquence U qui commence par le mot $w = GAGCACAGTGGGAAGGG$. Chaque nœud est étiqueté par son ordre d'insertion par souci de lisibilité. La 1^{ère} lettre de cette séquence est $U_1 = G$. Le premier nœud est placé dans l'arbre au niveau 1, sous la racine, sous la 3^{ème} branche car G est la 3^{ème} lettre de l'alphabet ordonné.

On insère ensuite le mot $U_1U_2 = GA$. Au niveau 1, sous la 1^{ère} branche car A est la 1^{ère} lettre de l'alphabet ordonné et le nœud correspondant est vide. On peut donc y insérer U_1U_2 . On itère ainsi successivement en ajoutant les nœuds correspondant aux mots $U_1\dots U_n$, pour $n \in \{3, \dots, 15\}$.

Il reste à placer le dernier nœud, c'est-à-dire celui qui correspond au mot w tout entier. On regarde la dernière lettre $U_{16} = G$. Au niveau 1, sous la 3^{ème} branche, le nœud est occupé par le premier mot. On descend sous ce nœud, dans la 3^{ème} branche puisque $U_{15} = G$. Le nœud est encore occupé, par le 11^{ème} mot. Comme $U_{14} = G$, on descend au niveau 3 sous la 3^{ème} branche. La place est libre, on peut y insérer le 16^{ème} mot.

Relation avec l'arbre digital de recherche

Un arbre digital de recherche est construit sur une suite de mots dont les lettres sont à valeurs dans un alphabet ordonné et fini de taille m . Cet arbre est composé d'autant de nœuds que de mots insérés. Chaque nœud de l'arbre possède m branches correspondant à l'alphabet ordonné. Le principe de construction est le suivant. L'arbre \mathcal{T}_n pousse en insérant un mot $W(n+1)$ sur un nœud étiqueté $W(n+1)$ pour former l'arbre \mathcal{T}_{n+1} . Ainsi les sous-arbres (\mathcal{T}_n) sont emboîtés. Pour ajouter le nœud étiqueté $W(n+1) \stackrel{\text{def}}{=} w_1w_2\dots$ on commence par regarder la première lettre w_1 . On cherche à l'insérer sous la racine, dans la branche correspondant à w_1 . Si le nœud n'est pas dans \mathcal{T}_n alors on y insère $W(n+1)$. Sinon, on descend d'un niveau dans la branche correspondant à w_2 . Si le nœud n'est pas dans \mathcal{T}_n on y insère $W(n+1)$, sinon on descend à nouveau et on itère le procédé jusqu'à

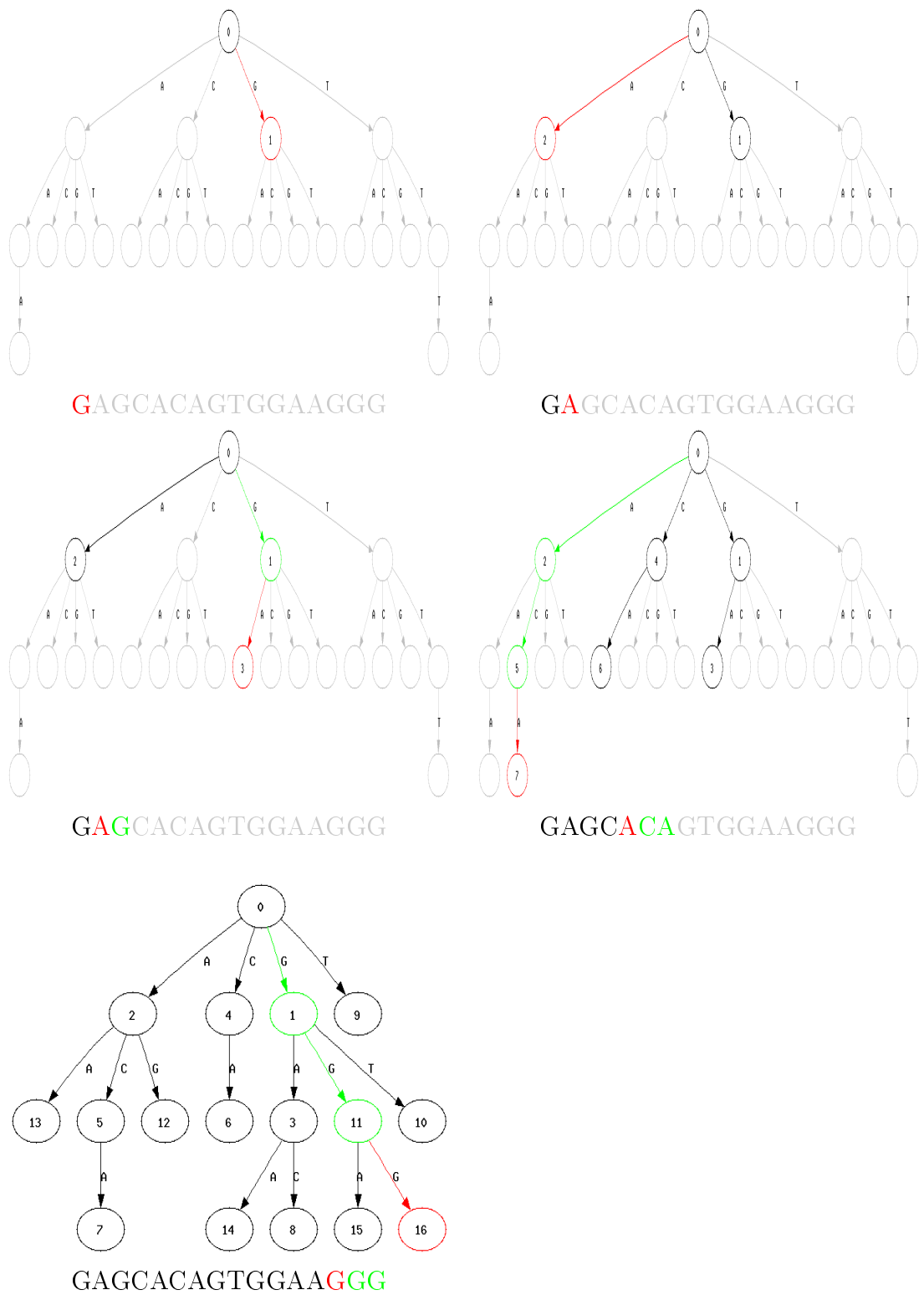


FIG. 3.1: Étapes successives de construction de l'arbre-CGR représentant la séquence GAGCACAGTGGAAAGGG.

pouvoir insérer $W(n+1)$ sur un nœud non présent dans \mathcal{T}_n . En particulier on remarque que les mots insérés ne se trouvent pas uniquement dans les feuilles de l'arbre.

Proposition 3.2.1. *L'arbre-CGR d'une séquence aléatoire $U = U_1U_2\dots$ est un arbre digital de recherche (DST, de l'anglais Digital Search Tree) obtenu par l'insertion successive dans un arbre quaternaire des préfixes retournés de la séquence U :*

$$\begin{aligned} W(1) &= U_1, \\ W(2) &= U_2U_1, \\ &\vdots \\ W(n) &= U_nU_{n-1}\dots U_1, \\ &\vdots \end{aligned} \tag{3.1}$$

Démonstration La preuve est évidente. ■

Les mots insérés sont donc fortement dépendants, contrairement aux DST classiques où les séquences insérées sont indépendantes les unes des autres. Dans ce chapitre, nous nous intéressons aux propriétés asymptotiques de cet arbre, construit à partir de séquences *markoviennes*. En particulier, nous démontrons un résultat de convergence presque sûre pour les longueurs des branches, ainsi qu'une propriété de convergence en probabilité pour la profondeur d'insertion.

Plus précisément, nous montrons que la profondeur d'une séquence aléatoire donnée dans l'arbre-CGR est asymptotiquement équivalente à la profondeur d'un DST construit à partir d'insertions indépendantes. De même, les longueurs des branches critiques sont asymptotiquement équivalentes pour ces deux schémas.

Forme de l'arbre et CGR

La donnée d'un arbre sans étiquette est équivalente à une liste de mots présents dans la séquence, sans tenir compte de l'ordre des occurrences. Plus précisément, on peut associer à un arbre sans étiquette un nuage de points dans le carré de la façon suivante. Chaque nœud de l'arbre est en bijection avec un mot $w = w_1\dots w_d$ et on lui associe le point X_w au centre du carré Sw correspondant, tel que, d'après (1.5),

$$X_w \stackrel{\text{def}}{=} \sum_{k=1}^d \frac{\ell_{w_k}}{2^{d-k+1}} + \frac{X_0}{2^d}. \tag{3.2}$$

Pour une occurrence d'un mot w , le *recentrage* dans l'équation (3.2) agit en quelque sorte comme un effaceur du passé, i.e. des symboles apparus avant w . Sur l'exemple de la Figure 3.2, le nuage de points dans le carré unité donne la forme de l'arbre pour le mot GAGCACAGTGAAGGG. De plus, la Figure 3.3 permet de comparer qualitativement

la représentation CGR avec le nuage de points associé à la forme de l'arbre-CGR d'une séquence d'ADN d'*Homo Sapiens* de longueur 400 000.

L'arbre-CGR est en bijection avec la séquence que l'on y insère. Au contraire, un arbre sans étiquettes peut représenter plusieurs séquences. On peut penser au simple exemple des mots *ATCG* et *CGAT*. Les deux arbres-CGR ont la même forme : le niveau 1 est rempli entièrement et tous les autres niveaux sont vides. Seules les étiquettes permettent de retrouver les séquences représentées. La forme de l'arbre indique seulement que les quatre lettres *A, C, G, T* ont été rencontrées. On *oublie* ce qui précédait l'occurrence de chaque lettre.

Remarque 3.2.2. Visuellement, il n'est pas facile de comparer ces deux représentations. L'aspect fractal dû à l'absence des mêmes mots dans les deux figures provoque une similarité. Cependant, nous nous sommes attachés, dans les chapitres précédents, à mettre en évidence l'information supplémentaire apportée par la donnée des points de la CGR, où chaque point contient toute l'histoire de la séquence, par rapport au comptage de mots. Finalement, la ressemblance des deux représentations de la Figure 3.3 est un leurre.

Bien que l'idée de construction de ces arbres quaternaires soit librement inspirée de la CGR, ils peuvent être obtenus directement à partir de la séquence. Cependant, cette construction est motivée par la volonté de mesurer des quantités statistiques cachées dans la séquence mais visibles et *révélées* sur l'arbre, afin de dégager de nouvelles caractéristiques pour une loi de génération fixée.

3.2.2 Liens avec les arbres des suffixes

L'arbre-CGR donne des indications sur les répétitions de mots dans la séquence étudiée. On se donne un nœud n_w de l'arbre, en bijection avec le mot $w = w_1 \dots w_d$. Chaque mot $W(i)$ situé dans le sous-arbre issu de n_w a pour suffixe w : il représente donc une occurrence de w dans la séquence. C'est d'ailleurs pour cette raison que nous avons *retourné* les séquences avant insertion dans l'arbre digital.

Cependant, toutes les occurrences de w ne sont pas représentées par des nœuds dans le sous-arbre issu de n_w . Par exemple, si la première occurrence de w dans la séquence U coïncide avec celle de l'un de ses suffixes $w_k \dots w_d$, la séquence sera insérée à un niveau $\leq d - k + 1$, et non sous le nœud n_w . Ainsi, le nombre d'occurrences de w est compris entre le nombre de nœuds N_w contenus dans le sous-arbre issu de w et $N_w + d - 1$.

La visualisation des répétitions de sous-mots dans l'arbre-CGR suggère un rapprochement avec *l'arbre des suffixes*. Ces arbres ont été introduits par Weiner [92] pour accélérer les opérations de recherche de motifs. Ils se construisent sur le même schéma récursif que les *tries*, à partir de l'ensemble des suffixes d'une séquence donnée. Le *trie* utilise une règle de construction récursive qui sépare les mots selon leurs préfixes. On stoppe la procédure dès que tous les mots sont distingués. Le nombre de feuilles du *trie* est égal au nombre de mots dans l'ensemble que l'on insère. Contrairement au DST, le *trie* est construit sur un ensemble et non sur une suite ordonnée de mots. De plus, on

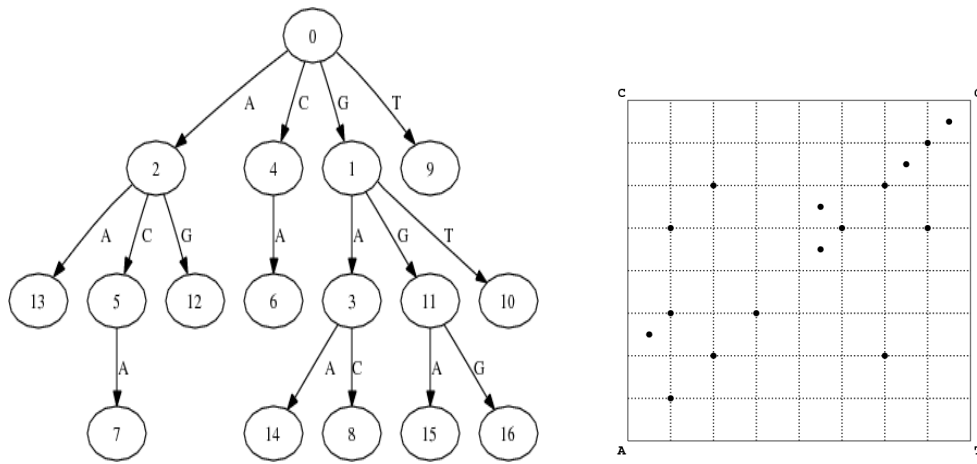


FIG. 3.2: Représentation de 16 nucléotides d'une séquence de *Mus Musculus* GAGCA-CAGTGGGAAGGG dans l'arbre-CGR (à gauche) et dans le carré unité (à droite). La grille en pointillés correspond aux ensembles S_w pour des mots w de longueur 3 (cf. Figure 1.1).

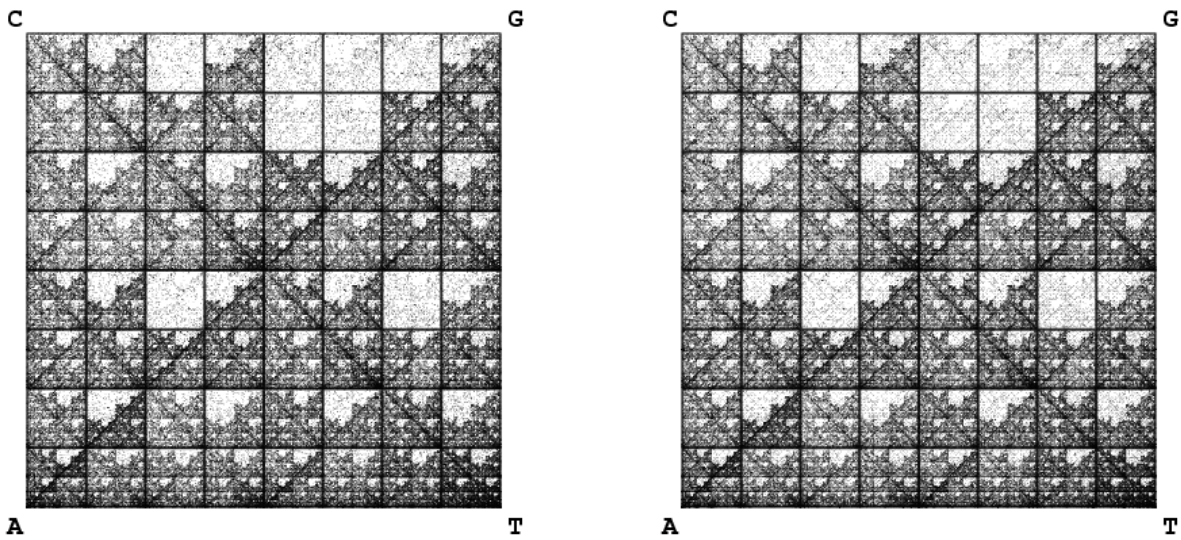


FIG. 3.3: Chaos Game Representation (sur la gauche) et représentation du nuage de points associé à l'arbre sans étiquettes (sur la droite) des 400 000 premiers nucléotides du Chromosome 2 d'*Homo Sapiens*.

effectue une comparaison de symboles, et non de mots, en chaque nœud interne. Pour un arbre des suffixes, on place dans un trie la suite des séquences complémentaires de celles que nous insérons dans l'arbre-CGR et définies en (3.1), i.e. les séquences

$$\begin{aligned}
 &U_1U_2\dots, \\
 &U_2U_3\dots, \\
 &\vdots \\
 &U_nU_{n+1}\dots, \\
 &\vdots
 \end{aligned}$$

La recherche d'un mot w s'effectue efficacement dans un arbre des suffixes. Il suffit de descendre dans la branche correspondant au mot w ; si le nœud est présent, alors w est dans la séquence.

L'arbre-CGR permet aussi de rechercher rapidement des mots dans la séquence. De même, si le nœud correspondant à w est présent, alors w est dans la séquence; toutefois, on ne connaît que le nombre maximum de ses occurrences. D'autre part, au contraire de l'arbre des suffixes, l'absence d'un nœud dans l'arbre-CGR ne signifie pas que le mot était absent de la séquence.

Les arbres des suffixes possèdent de nombreuses applications en informatique. Ils sont notamment à la base de l'algorithme de compression LZ'77 (Ziv et Lempel [94]). Ils ont été introduits en bioinformatique par Martinez [67] pour repérer des répétitions de mots. On utilise également les arbres des suffixes pour déterminer le plus long motif commun à deux séquences et ainsi faire de l'alignement de séquences, ou encore pour trouver le plus long palindrome d'une chaîne.

Dans sa thèse, Fayolle [35] établit le comportement asymptotique de l'espérance et de la variance des paramètres de taille, longueur de cheminement et profondeur typique, pour des arbres des suffixes construits à partir de séquences markoviennes. Il montre que la différence de comportement entre les arbres des suffixes et les tries sont asymptotiquement faibles.

De même, le comportement asymptotique (au 1^{er} ordre) d'arbres binaires de recherche (a.b.r.) construits à partir de séquences générées par une source indépendante, identiquement distribuée et uniforme est identique à celui d'a.b.r. construits à partir de la suite des suffixes d'une seule séquence (Devroye et Neininger [25]). Un a.b.r. est un arbre binaire dont chaque nœud interne est muni d'une étiquette, de sorte que toutes les étiquettes du sous-arbre droit soient plus *grandes* que toutes celles du sous-arbre gauche.

Nous verrons dans ce chapitre que la différence de comportement asymptotique entre les DST classiques et les arbres-CGR, issus de séquences markoviennes, n'est pas visible au 1^{er} ordre.

3.2.3 Occurrences de mots et recouvrements

La difficulté principale dans l'étude des propriétés asymptotiques des arbres-CGR provient de la dépendance des mots à insérer et des structures potentiellement auto-recouvrantes des mots. Les preuves utilisées dans ce travail font appel aux nombreuses études donnant des résultats sur la loi des positions d'occurrences d'un mot le long d'une séquence, ou encore sur le nombre d'occurrences d'un motif dans une séquence de longueur donnée. Blom et Thorburn [11] déterminent la fonction génératrice de ces lois dans le cas de séquences i.i.d., à partir d'une relation de récurrence sur les probabilités. Ce résultat est généralisé par Robin et Daudin [78] dans le cas d'une séquence markovienne d'ordre 1. De nombreuses études s'appuient sur les fonctions génératrices, par exemple Régnier [76], Reinert et al. [77], Stefanov et Pakes [87]. Cependant, il existe d'autres approches : l'une des techniques les plus générales est la *Markov chain embedding method* introduite par Fu [37] et développée par exemple par Fu et Koutras [38], Koutras [56]. L'approche martingale (voir par exemple Li [62], Gerber et Li [39], Williams [93]) est une alternative pour résoudre les problèmes liés au jeu de Penney [71]. Les deux approches sont comparées dans Pozdnyakov et al. [75].

Jeu de Penney

Le jeu de Penney repose sur l'étude des occurrences de séquences binaires données à l'intérieur d'une suite de tirages de pile ou face, à savoir une suite de variables aléatoires i.i.d. suivant la loi de Bernoulli de paramètre p . Le jeu consiste à faire jouer deux mots binaires A et B (non nécessairement de même longueur) l'un contre l'autre jusqu'à l'instant d'apparition du premier d'entre eux. Le mot gagnant est le mot qui apparaît en premier. Selon le choix des mots A et B , un joueur a plus de chance de gagner (même dans le cas équiprobable où $p = 1/2$), ou les deux joueurs peuvent gagner avec la même probabilité. Le but du jeu est de calculer la durée moyenne d'une partie ainsi que la probabilité de gagner de chacun des joueurs. Il est amusant de voir que deux mots équiprobables n'ont pas la même probabilité de gagner. Celle-ci dépend de la structure de recouvrement des mots. Il existe de nombreuses généralisations du jeu de Penney : par exemple, plus de deux joueurs ou bien des séquences à valeurs dans un alphabet de taille ≥ 2 . Citons à ce propos les travaux de Li [62], Stark [86] ou encore Knuth [55].

Bien que ces résultats ne soient pas directement utilisés dans nos preuves, ils ont enrichi notre intuition concernant les problèmes liés au recouvrement.

Taille de la plus longue répétition

Un lemme intermédiaire utilisé dans les preuves permet de déduire des propriétés asymptotiques sur la longueur des plus grandes répétitions de lettres. Dans le cas des séquences i.i.d. et symétriques, Erdős et Révész [32] établissent des résultats presque sûrs. Ces résultats sont étendus aux chaînes de Markov dans Samarova [83] tandis que

Gordon et al. [42] montrent que le comportement probabiliste de la longueur de la plus longue répétition peut être approché par un maximum de variables exponentielles i.i.d.

3.2.4 État de l'art sur les DST

Plusieurs résultats sont connus (voir chap. 6 dans Mahmoud [66]) sur la hauteur, la profondeur d'insertion et le profil, pour des DST construits sur des suites de séquences *indépendantes*, toutes de même loi. Rappelons que, dans notre cas, les mots successivement insérés sont fortement dépendants les uns des autres car, si $i \leq n$, $W(i)$ est alors un suffixe de $W(n)$.

Dans le cas du modèle de Bernoulli, les arbres sont binaires et les séquences insérées sont indépendantes les unes des autres ; de plus, chaque séquence est une suite de variables aléatoires indépendantes et de même loi de Bernoulli $\mathcal{B}(1/2)$. On trouve de nombreux résultats sur ces objets dans Mahmoud [66].

Dans Aldous et Shields [2], les DST sont un cas particulier des arbres considérés. On note $\partial\mathcal{T}_n$ la frontière de l'arbre \mathcal{T}_n , c'est-à-dire l'ensemble des nœuds qui ne sont pas dans \mathcal{T}_n mais dont les ancêtres sont dans \mathcal{T}_n . Les arbres croissent en sélectionnant un nœud de $\partial\mathcal{T}_n$ et en l'ajoutant à \mathcal{T}_n pour former \mathcal{T}_{n+1} . La probabilité d'ajouter un nœud α de $\partial\mathcal{T}_n$ est alors

$$\frac{c^{-h(\alpha)}}{\sum_{v \in \partial\mathcal{T}_n} c^{-h(v)}},$$

où $h(v)$ désigne la hauteur du nœud v et c est un paramètre vérifiant $1 \leq c \leq 2$. Pour $c = 2$, on retrouve les DST symétriques. Les résultats concernent le profil de l'arbre et sa hauteur H_n . En particulier,

$$H_n - \frac{\log n}{\log 2} \xrightarrow[n \rightarrow \infty]{\text{P}} 0.$$

La méthode repose sur le plongement en temps continu. Barlow et al. [6] généralisent ces travaux au cas où $c < 1$.

Drmotá [27] montre que la hauteur des arbres digitaux de recherche est *concentrée* au sens où $\mathbb{E}[|H_n - \mathbb{E}(H_n)|^L]$ est asymptotiquement bornée pour tout $L > 0$ quand n est grand. La méthode est analytique et utilise la résolution de l'équation récurrente différentielle

$$G'_{n+1}(x) = G_n(x/2)^2, \quad n \geq 1,$$

où G_n est la fonction génératrice des $\mathbb{P}(H_k \leq n)$ définie par

$$G_n(x) = \sum_{k \geq 0} \mathbb{P}(H_k \leq n) \frac{x^k}{k!}.$$

Dans le cas où les séquences à insérer sont indépendantes, mais à valeurs dans un alphabet de m lettres et émises par des sources à *faible dépendance*, Pittel [74] obtient des

résultats de convergence sur la profondeur d'insertion et sur la hauteur. La séquence U générée par la source satisfait la condition suivante : on note \mathcal{F}_a^b la σ -algèbre engendrée par U_a, \dots, U_b pour $1 \leq a \leq b$. Il existe deux constantes c_1 et c_2 et un entier $b_0 > 0$ tels que, pour $1 \leq a < a + b_0 < b$, on a

$$c_1 P(A)P(B) \leq P(A \cap B) \leq c_2 P(A)P(B),$$

où $A \in \mathcal{F}_1^a$ et $B \in \mathcal{F}_{a+b_0}^b$. Cette condition entraîne que la séquence U est fortement mélangée (voir par exemple Billingsley [10]). En particulier, une source i.i.d. (symétrique comme biaisée) ou markovienne, générant des chaînes irréductibles apériodiques, est un cas particulier de telles sources.

Exceptée l'indépendance des séquences, le travail de Pittel semble être le plus proche du nôtre et plusieurs parties de démonstrations en sont inspirées.

Soient D_n la profondeur d'insertion, ℓ_n la longueur des branches les plus courtes et \mathcal{L}_n la longueur des branches les plus longues.

Théorème 3.2.3 (Pittel, 1985).

$$\frac{\ell_n}{\log n} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \frac{1}{h_+} \quad \text{et} \quad \frac{\mathcal{L}_n}{\log n} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \frac{1}{h_-},$$

où h_+ et h_- sont des constantes dépendant de la loi de génération de la source que nous définissons en Section 3.3.1.

Théorème 3.2.4 (Pittel, 1985).

$$\frac{D_n}{\log n} \xrightarrow[n \rightarrow \infty]{\text{P}} \frac{1}{h},$$

où la constante h est l'entropie de cette source.

Les arbres-CGR ne sont pas construits à partir de suites de mots indépendants, mais à partir de séquences fortement dépendantes les unes des autres. Nous montrons dans ce chapitre que les deux théorèmes de Pittel restent vrais dans ce cas. La difficulté principale est de comprendre comment pousse l'arbre-CGR et comment on *reconnaît* une séquence dans l'arbre.

3.3 Convergence presque sûre de branches critiques

3.3.1 Préambule

Dans tout ce chapitre, on supposera que la séquence $U = U_1 U_2 \dots U_n \dots$ forme une chaîne de Markov d'ordre 1 à espace d'états fini $\mathcal{A} = \{A, C, G, T\}$, irréductible, apériodique, stationnaire, de matrice de transition Q et de mesure invariante p .

Pour une séquence infinie fixée déterministe s , on note $s^{(n)}$ le mot constitué des n premières lettres de s , c'est-à-dire $s^{(n)} = s_1 \dots s_n$, où s_i désigne la $i^{\text{ème}}$ lettre de s . La mesure

p est étendue aux mots $s^{(n)}$ retournés en posant $p(s^{(n)}) \stackrel{\text{def}}{=} \mathbb{P}(U_1 = s_n, \dots, U_n = s_1)$. On inverse la séquence à cause de la construction de l'arbre-CGR et de la Proposition 3.2.1. Les mots à insérer sont les séquences retournées définies dans l'équation (3.1). Dans le cas où U est i.i.d., la probabilité $p(s^{(n)})$ est simplement le produit des probabilités des lettres constituant le mot $s^{(n)}$. Dans le modèle markovien, on a

$$p(s^{(n)}) = p(s_n)Q(s_n, s_{n-1}) \dots Q(s_2, s_1).$$

On définit également les constantes

$$\begin{aligned} h_+ &\stackrel{\text{def}}{=} \lim_{n \rightarrow +\infty} \frac{1}{n} \max \left\{ \log \left(\frac{1}{p(s^{(n)})} \right) \mid p(s^{(n)}) > 0 \right\}, \\ h_- &\stackrel{\text{def}}{=} \lim_{n \rightarrow +\infty} \frac{1}{n} \min \left\{ \log \left(\frac{1}{p(s^{(n)})} \right) \mid p(s^{(n)}) > 0 \right\}, \\ h &\stackrel{\text{def}}{=} \lim_{n \rightarrow +\infty} \frac{1}{n} \mathbb{E} \left[\log \left(\frac{1}{p(U^{(n)})} \right) \right], \end{aligned}$$

où le max et le min sont pris sur tous les mots $s^{(n)}$ avec $p(s^{(n)}) > 0$. Dans le cas particulier où U est i.i.d., la constante h_- est liée à p_{max} , la plus grande des 4 probabilités, par la relation $h_- = -\log p_{max}$, et $h_+ = -\log p_{min}$ où p_{min} est la plus petite probabilité. Pittel [74] montre que ces limites sont bien définies (dans son cadre plus général de faible dépendance), grâce à un argument de sous-additivité. D'autre part il prouve que les deux limites h_+ et h_- sont atteintes. On note alors s_+ et s_- des séquences infinies vérifiant

$$h_+ = \lim_{n \rightarrow \infty} \frac{1}{n} \log \left(\frac{1}{p(s_+^{(n)})} \right) \quad \text{et} \quad h_- = \lim_{n \rightarrow \infty} \frac{1}{n} \log \left(\frac{1}{p(s_-^{(n)})} \right). \quad (3.3)$$

Dans notre cadre markovien fini, ces constantes ont les interprétations suivantes. On associe à chaque cycle simple c , un poids dépendant de la matrice de transition Q par la relation

$$q(c) \stackrel{\text{def}}{=} \prod_{j=1}^{|c|} Q(c_j, c_{j+1}).$$

Les constantes h_+ et h_- s'écrivent alors

$$\begin{aligned} e^{-h_+} &= \min \left\{ q(c)^{1/c}; c \text{ cycle simple, } q(c) \neq 0 \right\}, \\ e^{-h_-} &= \max \left\{ q(c)^{1/c}; c \text{ cycle simple, } q(c) \neq 0 \right\}. \end{aligned}$$

Pour $n \geq 1$, soit $\mathcal{T}_n \stackrel{\text{def}}{=} \mathcal{T}_n(U)$ l'arbre fini à n nœuds (sans compter la racine), construit à partir des n premières séquences $W(1), \dots, W(n)$. L'arbre ne contenant que la racine est noté \mathcal{T}_0 .

Il est clair, par construction, que les arbres \mathcal{T}_n sont emboîtés, $\mathcal{T}_0 \subset \mathcal{T}_1 \subset \dots \mathcal{T}_n \subset \dots$

On introduit les variables fondamentales suivantes :

- ℓ_n , la longueur du plus court chemin de la racine à une feuille de l'arbre \mathcal{T}_n ;
- \mathcal{L}_n , la longueur du chemin le plus long de la racine à une feuille de l'arbre \mathcal{T}_n ;
- D_n , la profondeur d'insertion de $W(n)$ dans l'arbre \mathcal{T}_{n-1} (pour créer \mathcal{T}_n) ;
- M_n , la longueur d'un chemin de l'arbre \mathcal{T}_n , choisi aléatoirement et uniformément parmi les n chemins possibles.

Notons que \mathcal{L}_n représente la hauteur de l'arbre. Quant à ℓ_n , elle donne des renseignements sur le niveau de saturation.

Les variables aléatoires définies ci-dessous jouent un rôle clé dans les preuves. Pour bien fixer le cadre, on rappelle que s est une donnée déterministe et que l'aléa n'est engendré que par la séquence U , c'est-à-dire par la construction des arbres \mathcal{T}_n .

$$X_n(s) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{si } s_1 \text{ n'est pas dans } \mathcal{T}_n \\ \max\{k \geq 1 \mid \text{le mot } s^{(k)} \text{ est déjà inséré dans } \mathcal{T}_n\} \end{cases}$$

$$T_k(s) \stackrel{\text{def}}{=} \min\{n \geq 1 \mid X_n(s) = k\}.$$

$T_k(s)$ désigne ainsi la taille du premier arbre où $s^{(k)}$ est inséré. On peut noter que $T_0(s) = 0$. Ces deux variables sont en dualité au sens où

$$\{X_n(s) \geq k\} = \{T_k(s) \leq n\}. \quad (3.4)$$

Ici on obtient donc $\{T_k(s) = n\} \subset \{X_n(s) = k\}$. La variable $X_n(s)$ désigne la longueur de la branche correspondant à s dans l'arbre \mathcal{T}_n . Il est important de noter que dans tout arbre \mathcal{T}_n , la lecture des préfixes $s^{(k)}$ doit être faite *de haut en bas*, c'est-à-dire de la racine vers les feuilles.

Reprenons l'exemple de construction de la Figure 3.1. Choisissons arbitrairement une séquence infinie déterministe s telle que $s^{(3)} = \text{ACA}$. Pour l'arbre-CGR d'une séquence infinie commençant par GAGCACAGTGGGAAGGG, on a

$$X_0(s) = X_1(s) = 0, \quad X_2(s) = X_3(s) = X_4(s) = 1, \quad X_5(s) = X_6(s) = 2$$

et, pour $n \geq 7$, $X_n(s) = 3$. On peut aussi indiquer les valeurs de la variables $T_k(s)$:

$$T_1(s) = 2, \quad T_2(s) = 5, \quad T_3(s) = 7.$$

La variable $T_k(s)$ admet la décomposition

$$T_k(s) = \sum_{r=1}^k Z_r(s), \quad (3.5)$$

où $Z_r(s) \stackrel{\text{def}}{=} T_r(s) - T_{r-1}(s)$ est le nombre de symboles à lire pour que la longueur de la branche décrivant la séquence s augmente de 1. Du point de vue de la séquence, c'est aussi le temps d'attente n de la première occurrence de $s^{(r)}$ dans la séquence retournée

$$\dots U_{n+T_{r-1}(s)} U_{n-1+T_{r-1}(s)} \dots U_{1+T_{r-1}(s)} s^{(r-1)}.$$

$U_{3+T_5(s)}$	$U_{2+T_5(s)}$	$U_{1+T_5(s)}$	s_1	s_2	s_3	s_4	s_5	
s_1	s_2	s_3	s_4	s_5	s_6			

FIG. 3.4: Importance de la structure de recouvrement dans la définition de la variable aléatoire $Z_r(s)$. Ici, à titre d'exemple, la longueur du mot considéré est $r = 6$. Une occurrence de $s^{(6)}$ peut exister en $U_{3+T_5(s)}$ seulement si $s_1s_2s_3 = s_4s_5s_6$.

La variable aléatoire $Z_r(s)$ est donc aussi définie par

$$Z_r(s) \stackrel{\text{def}}{=} \min\{n \geq 1 \mid U_{n+T_{r-1}(s)} \cdots U_{n+T_{r-1}(s)-r+1} = s_1 \cdots s_r\}.$$

Il est assez facile de voir que les variables aléatoires $Z_r(s)$ sont indépendantes (par un raisonnement analogue à celui qui permet d'établir l'indépendance des excursions d'une chaîne de Markov en dehors d'un ensemble donné).

On introduit également la variable $Y_r(s)$ représentant le temps d'attente n de la première occurrence du mot $s^{(r)}$ dans la séquence

$$\cdots U_{n+T_{r-1}(s)} U_{n-1+T_{r-1}(s)} \cdots U_{1+T_{r-1}(s)},$$

c'est-à-dire

$$Y_r(s) \stackrel{\text{def}}{=} \min\{n \geq r \mid U_{n+T_{r-1}(s)} \cdots U_{n+T_{r-1}(s)-r+1} = s_1 \cdots s_r\}.$$

Par définition on a immédiatement $Z_r(s) \leq Y_r(s)$. Si l'occurrence du mot $s^{(r)}$ se produit avant $T_{r-1}(s) + r$, c'est qu'il existe une structure de recouvrement entre les préfixes de $s^{(r-1)}$ et les suffixes de $s^{(r)}$. La Figure 3.4 illustre cette assertion sur un exemple pour $r = 6$ et dans le cas où $s_1s_2s_3 = s_4s_5s_6$.

Plus précisément les variables $Z_r(s)$ et $Y_r(s)$ sont liées par la relation

$$Z_r(s) = \mathbb{1}_{\{Z_r(s) < r\}} Z_r(s) + \mathbb{1}_{\{Z_r(s) \geq r\}} Y_r(s).$$

Comme la séquence est stationnaire, la loi de $Y_r(s)$ est celle du temps d'attente de la première occurrence du mot $s_r \cdots s_1$ dans une réalisation d'une chaîne de Markov de matrice de transition Q évoluant en régime stationnaire. La fonction génératrice $\Phi(s^{(r)}, t) \stackrel{\text{def}}{=} \mathbb{E}[t^{Y_r(s)}]$ de $Y_r(s)$ est donnée par Robin et Daudin [78] sous la forme

$$\Phi(s^{(r)}, t) = \left(\gamma_r(t) + (1-t)\delta_r(t^{-1}) \right)^{-1}, \quad (3.6)$$

où les fonctions γ_r et δ_r sont respectivement définies par

$$\gamma_r(t) \stackrel{\text{def}}{=} \frac{1-t}{tp(s_r)} \sum_{m \geq 1} Q^m(s_1, s_r) t^m, \quad \delta_r(t^{-1}) \stackrel{\text{def}}{=} \sum_{m=1}^r \frac{\mathbb{1}_{\{s_r \dots s_{r-m+1} = s_m \dots s_1\}}}{t^m p(s^{(m)})}, \quad (3.7)$$

et $Q^m(u, v)$ est la probabilité de transition de u vers v en m étapes pour une chaîne de Markov de matrice de transition Q .

Remarque 3.3.1. Dans le cas particulier où la séquence de nucléotides U est supposée indépendante et identiquement distribuée selon la loi non dégénérée (p_A, p_C, p_G, p_T) , la probabilité de transition $Q^m(s_1, s_r)$ est égale à $p(s_r)$ et par conséquent $\gamma_r(t) = 1$.

Remarque 3.3.2. La clé des preuves réside dans l'expression des fonctions génératrices de $Y_r(s)$ et $Z_r(s)$. Dans un cadre de dépendance faible des sources comme celui de Pittel, il n'est pas possible d'écrire explicitement ces fonctions. C'est pourquoi nous nous sommes « restreints » à une hypothèse markovienne pour la source des lettres de la séquence U à la base de la construction de l'arbre-CGR.

Proposition 3.3.3.

(i) La fonction $\Phi(s^{(r)}, t)$ est au moins définie sur l'intervalle réel $[0, 1 + \kappa p(s^{(r)})[$, où κ est une constante positive indépendante de r et s .

(ii) Soit γ la plus grande valeur propre différente de 1 de la matrice de transition Q . Pour tout t vérifiant $|t| < |\gamma|^{-1}$

$$|\gamma_r(t) - 1| \leq \frac{|1-t|}{1-|\gamma t|} \kappa',$$

où κ' est une constante indépendante de r et s .

Démonstration La preuve de la Proposition 3.3.3 est donnée en Section 3.7. ■

Le théorème suivant établit le comportement asymptotique des variables ℓ_n et \mathcal{L}_n de l'arbre-CGR.

Théorème 3.3.4 (Convergence p.s. des branches critiques).

$$\frac{\ell_n}{\log n} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \frac{1}{h_+} \quad \text{et} \quad \frac{\mathcal{L}_n}{\log n} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \frac{1}{h_-}.$$

Remarque 3.3.5. On retrouve bien les mêmes constantes limites que dans le théorème 3.2.3 de Pittel, où les DST sont construits à partir de séquences indépendantes.

Par définition, $X_n(s)$ est la variable représentant la longueur de la branche correspondant à s dans l'arbre \mathcal{T}_n . Ainsi les longueurs ℓ_n et \mathcal{L}_n peuvent-elles naturellement s'exprimer en fonction de X_n par les relations

$$\ell_n = \min_{s \in \mathcal{A}^{\mathbb{N}}} X_n(s) \quad \text{et} \quad \mathcal{L}_n = \max_{s \in \mathcal{A}^{\mathbb{N}}} X_n(s). \quad (3.8)$$

Le lemme clé suivant établit un résultat asymptotique sur $X_n(s)$, sous certaines conditions sur la séquence déterministe s . On déduira de ce lemme des résultats asymptotiques sur les longueurs des branches critiques.

3.3.2 Lemme préliminaire

Lemme 3.3.6. *Supposons pour une séquence infinie s qu'on puisse définir la limite*

$$\lim_{n \rightarrow +\infty} \frac{1}{n} \log \left(\frac{1}{p(s^{(n)})} \right) = h(s) > 0. \quad (3.9)$$

Alors, on a la convergence

$$\frac{X_n(s)}{\log n} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \frac{1}{h(s)}.$$

Remarque 3.3.7. Soit $\underline{v} \stackrel{\text{def}}{=} vv\dots$ un mot infini constitué de répétitions d'une seule lettre v . La variable $X_n(\underline{v})$ désigne la longueur de la branche associée à \underline{v} dans \mathcal{T}_n . Pour une telle séquence (et uniquement pour une telle séquence) chaque variable aléatoire $Y_k(\underline{v})$ est égale à $T_k(\underline{v})$. En d'autres termes, l'insertion du nœud au niveau k dans la branche correspondante à \underline{v} coïncide avec la première occurrence du mot $\underline{v}^{(k)}$. En effet, lorsqu'une occurrence d'une répétition $\underline{v}^{(k)}$ apparaît dans la séquence en position n , il y avait obligatoirement une occurrence de $\underline{v}^{(k-1)}$ en $n-1$ et ainsi de suite. Par conséquent, $X_n(\underline{v})$ est la longueur de la plus longue répétition de la lettre v dans $U_1 \dots U_n$. Lorsque U est une séquence de lettres i.i.d., Petrov [73], Erdős et Révész [33], Erdős et Révész [32] montrent que

$$\frac{X_n(\underline{v})}{\log n} \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \frac{1}{\log p},$$

où $p \stackrel{\text{def}}{=} \mathbb{P}(U_n = v)$. Ce résultat de convergence est un cas particulier du lemme 3.3.6.

Preuve du lemme 3.3.6 Par des arguments de monotonie, il est suffisant de montrer

$$\frac{\log T_k(s)}{k} \xrightarrow[k \rightarrow \infty]{\text{p.s.}} h(s). \quad (3.10)$$

En effet, la suite $T_k(s)$ est croissante et tend presque sûrement vers l'infini. Donc, pour tout entier $n \geq 1$ assez grand, il existe $k \geq 1$ tel que

$$3 \leq T_k(s) \leq n < T_{k+1}(s),$$

ce qui implique

$$k \leq X_n(s) < k + 1.$$

D'où l'on tire

$$\left(\frac{k}{k+1} \right) \frac{k+1}{\log T_{k+1}(s)} < \frac{X_n(s)}{\log n} < \left(\frac{k+1}{k} \right) \frac{k}{\log T_k(s)}.$$

Par passage à la limite, le théorème des gendarmes entraîne immédiatement le lemme 3.3.6. Il suffit donc de prouver la convergence (3.10).

On commence par écrire

$$\frac{1}{k} \log T_k(s) = \frac{1}{k} \log \frac{T_k(s)}{\mathbb{E}[T_k(s)]} + \frac{1}{k} \log \mathbb{E}[T_k(s)].$$

La variable $T_k(s)$ est une somme de variables indépendantes grâce à l'équation (3.5). La variance et l'espérance de $Z_r(s)$ sont données par Robin et Daudin [78]

$$\mathbb{E}[Z_r(s)] = \frac{1}{p(s^{(r)})}, \quad \text{var}(Z_r(s)) \leq 4 \frac{r}{p(s^{(r)})^2}.$$

L'égalité

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log \mathbb{E}[T_k(s)] = h(s) \quad \text{p.s.}$$

est donc une conséquence immédiate du résultat élémentaire suivant : si (x_k) est une suite positive, alors

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log(x_k) = h > 0 \implies \lim_{k \rightarrow \infty} \frac{1}{k} \log \left(\sum_{r=1}^k x_r \right) = h.$$

Il reste à montrer que

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log \left(\frac{T_k(s)}{\mathbb{E}[T_k(s)]} \right) = 0 \quad \text{p.s.} \quad (3.11)$$

La preuve est basée sur le critère classique de convergence presque sûre suivant. Soit une suite de variables aléatoires (χ_n) vérifiant

$$\lim_{n \rightarrow \infty} \sum_{k \geq n} \mathbb{P}(|\chi_k - \chi| > \varepsilon) = 0, \quad \forall \varepsilon > 0.$$

Alors (χ_n) converge presque sûrement vers χ .

Ici, il s'agit de majorer

$$\mathbb{P} \left(\left| \frac{1}{k} \log \frac{T_k(s)}{\mathbb{E}[T_k(s)]} \right| > \varepsilon \right)$$

par le terme d'une série convergente. On a l'égalité évidente

$$\mathbb{P} \left(\left| \frac{1}{k} \log \frac{T_k(s)}{\mathbb{E}[T_k(s)]} \right| > \varepsilon \right) = \mathbb{P} \left(\frac{T_k(s)}{\mathbb{E}[T_k(s)]} > \exp(k\varepsilon) \right) + \mathbb{P} \left(\frac{T_k(s)}{\mathbb{E}[T_k(s)]} < \exp(-k\varepsilon) \right),$$

dans laquelle on applique l'inégalité de Bienaymé Tchebychev au premier terme du membre droit. On utilise aussi le résultat de convergence, valable pour tout réel $c > 1$,

$$\sum_{k=1}^n kc^k = \mathcal{O}(nc^n).$$

Il vient alors

$$\mathbb{P}\left(\frac{T_k(s)}{\mathbb{E}[T_k(s)]} < \exp(-k\varepsilon)\right) \leq \mathbb{E}[T_k(s)] \exp(-k\varepsilon) p(s^{(k)}) = \mathcal{O}(\exp(-k\varepsilon)),$$

on obtient

$$\mathbb{P}\left(\left|\frac{1}{k} \log \frac{T_k(s)}{\mathbb{E}[T_k(s)]}\right| > \varepsilon\right) = \mathcal{O}\left(k \exp(-2kh) + \exp(-k\varepsilon)\right).$$

Le critère de convergence énoncé plus haut conclut la preuve du lemme 3.3.6. ■

Remarque 3.3.8. On peut également donner une preuve utilisant la loi forte des grands nombres pour les martingales, en décomposant $T_k(s)$ sous la forme

$$T_k(s) = \sum_{r=1}^k \varepsilon_r(s) + \sum_{r=1}^k \mathbb{E}[Z_r(s)],$$

où $\varepsilon_r(s) \stackrel{\text{def}}{=} Z_r(s) - \mathbb{E}[Z_r(s)]$. En prenant le logarithme dans l'équation précédente, on a

$$\log T_k(s) = \log \mathbb{E}[T_k(s)] + \log\left(1 + \frac{M_k(s)}{\mathbb{E}[T_k(s)]}\right), \quad (3.12)$$

où la martingale $M_k(s)$ est définie par

$$M_k(s) \stackrel{\text{def}}{=} \sum_{k=1}^n \varepsilon_k.$$

La convergence du premier terme dans le membre droit de (3.12) a déjà été traitée dans la preuve du lemme 3.3.6. La suite (M_k) est une martingale de carré intégrable. Le processus croissant associé, noté $\langle M(s) \rangle_k$ peut se majorer, encore une fois en utilisant les résultats de Robin et Daudin [78] concernant $Z_r(s)$

$$\langle M(s) \rangle_k = \mathcal{O}\left(k \exp(2kh(s))\right).$$

On conclut en appliquant la loi des grands nombres pour martingales. Pour tout $\alpha > 0$, on a

$$M_k(s) = \mathcal{O}\left(\langle M(s) \rangle_k^{1/2} (\log \langle M(s) \rangle_k)^{(1+\alpha)/2}\right) \quad \text{p.s.}$$

Par conséquent,

$$\frac{M_k(s)}{\mathbb{E}[T_k(s)]} = \mathcal{O}\left(k^{1+\alpha/2}\right) \quad \text{p.s.},$$

ce qui achève également la preuve du lemme 3.3.6.

Démonstration du théorème 3.3.4 Elle est inspirée de celle de Pittel [74]. La définition donnée en (3.8) conduit clairement à

$$\ell_n \leq X_n(s_+) \quad \text{et} \quad \mathcal{L}_n \geq X_n(s_-).$$

Rappelons que s_+ et s_- ont été définies en (3.3). En particulier, on en déduit que s_+ et s_- satisfont l'hypothèse du lemme 3.3.6, qui donne

$$\limsup_{n \rightarrow \infty} \frac{\ell_n}{\log n} \leq \frac{1}{h_+}, \quad \liminf_{n \rightarrow \infty} \frac{\mathcal{L}_n}{\log n} \geq \frac{1}{h_-} \quad \text{p.s..}$$

Il reste à minorer la limite inférieure de $\ell_n/\log n$ et majorer la limite supérieure de $\mathcal{L}_n/\log n$.

3.3.3 Minoration de la limite inférieure de la longueur des plus courtes branches

Grâce à la relation (3.8) et à la dualité des variables $X_n(s)$ et $T_k(s)$ explicitée par (3.4), pour tout entier $k \geq 1$, on a l'inégalité

$$\mathbb{P}(\ell_n \leq k-1) \leq \sum_{s^{(k)} \in \mathcal{A}^k} \mathbb{P}(X_n(s) \leq k-1) \leq \sum_{s^{(k)} \in \mathcal{A}^k} \mathbb{P}(T_k(s) \geq n), \quad (3.13)$$

où les sommes sont prises sur l'ensemble des mots de longueur k . On sait d'après la Proposition 3.3.3 que, pour un réel $t \in [1, 1 + \kappa p(s^{(k)})[$, les fonctions génératrices $\Phi(s^{(r)}, t)$, $\forall r \leq k$ sont bien définies, tout comme les fonctions génératrices de $Z_r(s)$. De plus, comme les $(Z_r(s))$ sont indépendantes et vérifient l'inégalité $Z_r(s) \leq Y_r(s)$, chaque terme de la somme (3.13) satisfait, pour tout réel $t \geq 1$, l'inégalité

$$\mathbb{P}(T_k(s) \geq n) \leq t^{-n} \mathbb{E}[t^{T_k(s)}] \leq t^{-n} \prod_{r=1}^k \Phi(s^{(r)}, t).$$

En particulier, en majorant toutes les indicatrices de recouvrement par 1 dans (3.7), on déduit de (3.6), (3.7), et de la Proposition 3.3.3 que

$$\mathbb{P}(T_k(s) \geq n) \leq t^{-n} \prod_{r=1}^k \left(1 + (1-t) \left(\frac{\kappa'}{1-|\gamma|t} + \sum_{m=1}^r \frac{1}{t^m p(s^{(m)})} \right) \right)^{-1}.$$

D'après la définition de h_+ , on a pour r choisi suffisamment grand

$$p(s^{(r)}) > \alpha^r, \quad \alpha \stackrel{\text{def}}{=} \exp[-(1+\varepsilon^2)h_+] < 1,$$

avec $0 < \varepsilon < 1$. On peut alors trouver une constante $c > 0$ telle que

$$\mathbb{P}(T_k(s) \geq n) \leq ct^{-n} \prod_{r=1}^k \left(1 + (1-t) \left(\frac{1-(\alpha t)^{-r}}{(\alpha t - 1)} + \frac{\kappa'}{1-|\gamma|t} \right) \right)^{-1}.$$

En choisissant $t = 1 + \kappa\alpha^k$ où κ est la constante définie dans la Proposition 3.3.3, on obtient

$$\mathbb{P}(T_k(s) \geq n) \leq ct^{-n} \prod_{r=1}^k \left(1 - \alpha^{k-r} (a_r(k) + b_r(k))\right)^{-1}$$

avec

$$a_r(k) = \frac{\kappa(\alpha^r - (1 + \kappa\alpha^k)^{-r})}{\alpha(1 + \kappa\alpha^k) - 1},$$

$$b_r(k) = \frac{\alpha^r \kappa \kappa'}{1 - |\gamma|(1 + \kappa\alpha^k)}.$$

Comme $k \geq r$, si l'on fait tendre r vers l'infini, k tendra également vers l'infini. De l'inégalité élémentaire

$$x - \frac{x^2}{2} \leq \log(1 + x) \leq x$$

avec $x > 0$, en notant que $\alpha < 1$, on déduit les limites

$$\lim_{k \rightarrow \infty} \exp(-r \log(1 + \kappa\alpha^k)) = 1, \quad \lim_{k \rightarrow \infty} a_r(k) = \frac{1}{1 - \alpha}, \quad \lim_{k \rightarrow \infty} b_r(k) = 0.$$

Il existe donc deux nombres $\delta > 0$, et λ avec $0 < \alpha\lambda < 1$, tels que

$$\mathbb{P}(T_k(s) \geq n) \leq \delta t^{-n} \prod_{r=1}^k \left(1 - \lambda\alpha^{k-r}\right)^{-1}.$$

En utilisant l'inégalité évidente

$$\prod_{r=1}^k \left(1 - \lambda\alpha^{k-r}\right)^{-1} \leq \prod_{r=0}^{\infty} \left(1 - \lambda\alpha^r\right)^{-1} < \infty,$$

on conclut à l'existence d'un nombre $\tau > 0$ tel que

$$\mathbb{P}(T_k(s) \geq n) \leq \tau t^{-n}.$$

Choisissant maintenant $k = \lfloor (1 - \varepsilon) \log(n) / h_+ \rfloor + 1$, avec $0 < \varepsilon < 1$, on a pour $0 < x < 1$, l'inégalité

$$t^{-n} \leq \exp\left(-\frac{n\kappa\alpha^k}{2}\right) \leq \exp\left(-\frac{\kappa n^\theta}{2}\right),$$

où $\theta = \varepsilon(1 - \varepsilon + \varepsilon^2) > 0$. On déduit alors de (3.13) l'inégalité

$$\mathbb{P}(\ell_n \leq k - 1) \leq \tau 4^k \exp\left(-\frac{\kappa n^\theta}{2}\right),$$

dont le membre droit est le terme général d'une série absolument convergente vis à vis de la variable n . Le lemme de Borel-Cantelli montre alors, en faisant tendre ε vers zéro, que

$$\liminf_{n \rightarrow \infty} \frac{\ell_n}{\log n} \geq \frac{1}{h_+} \quad \text{p.s.}$$

Remarque 3.3.9. En majorant toutes les fonctions indicatrices de recouvrement par 1, c'est-à-dire en considérant le cas des mots constitués que d'une seule lettre, on a pu obtenir une majoration suffisamment fine pour ne pas avoir à distinguer entre les différentes « espèces » de recouvrements. En fait, la majoration de $\mathbb{P}(\ell_n \leq k)$ obtenue est grossière mais suffisante pour appliquer le lemme de Borel-Cantelli.

3.3.4 Majoration de la limite supérieure de la hauteur

Pour compléter la preuve, il reste à montrer que

$$\limsup_{n \rightarrow \infty} \frac{\mathcal{L}_n}{\log n} \leq \frac{1}{h_-} \quad \text{p.s.}$$

Grâce à la relation (3.8) et à la dualité des variables $X_n(s)$ et $T_k(s)$, avec un argument de monotonie analogue à celui utilisé dans la preuve du lemme 3.3.6, il suffit de prouver que

$$\liminf_{k \rightarrow \infty} \min_{s^{(k)} \in \mathcal{A}^k} \frac{\log T_k(s)}{k} \geq h_- \quad \text{p.s.}$$

Comme précédemment, on va appliquer le lemme de Borel-Cantelli en majorant par le terme d'une série convergente

$$\mathbb{P}\left(\min_{s^{(k)} \in \mathcal{A}^k} T_k(s) < \exp(kh_-(1 - \varepsilon))\right)$$

avec $0 < \varepsilon < 1$. En majorant brutalement la probabilité de l'union par la somme des probabilités, on obtient évidemment

$$\mathbb{P}\left(\min_{s^{(k)} \in \mathcal{A}^k} T_k(s) < \exp(kh_-(1 - \varepsilon))\right) \leq \sum_{s^{(k)} \in \mathcal{A}^k} \mathbb{P}(T_k(s) < n),$$

où $n \stackrel{\text{def}}{=} \exp(kh_-(1 - \varepsilon))$. Pour tout t dans $]0, 1[$, d'après l'inégalité de Markov

$$\mathbb{P}(T_k(s) < n) \leq t^{-n} \mathbb{E}[t^{T_k(s)}].$$

On déduit de la décomposition (3.5) et de l'indépendance des $Z_r(s)$ que

$$\mathbb{P}(T_k(s) < n) \leq t^{-n} \prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}]. \quad (3.14)$$

Afin de majorer le terme t^{-n} (avec $0 < t < 1$), on choisit $t \stackrel{\text{def}}{=} (1 + c/n)^{-1}$. La fonction génératrice de $Z_r(s)$ est donnée par Robin et Daudin [78] et dépend très fortement de la structure de recouvrement du mot $s^{(r)}$. Pour $0 < t < 1$, cette fonction est bien définie et

$$\mathbb{E}[t^{Z_r(s)}] = 1 - \frac{(1 - t)}{t^r p(s^{(r)}) (\gamma_r(t) + (1 - t)\delta_r(t^{-1}))}, \quad (3.15)$$

où $\gamma_r(t)$ et $\delta_r(t)$ sont données par (3.7). De plus, on déduit aisément de la Proposition 3.3.3 que, pour tout $0 < t < 1$,

$$\gamma_r(t) \leq 1 + \theta(1 - t) \quad \text{avec} \quad (1 - |\gamma|)\theta = \kappa. \quad (3.16)$$

Il s'agit d'étudier le comportement asymptotique de cette fonction génératrice. Par un simple changement de variables, il vient

$$\begin{aligned} t^r p(s^{(r)}) \delta_r(t^{-1}) &= \sum_{m=1}^r \mathbb{1}_{\{s_r \dots s_{r-m+1} = s_m \dots s_1\}} \frac{t^r p(s^{(r)})}{t^m p(s^{(m)})} \\ &= \sum_{m=1}^r \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} t^{m-1} \frac{p(s^{(r)})}{p(s^{(r-m+1)})} \\ &= \sum_{m=1}^r \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} t^{m-1} \frac{p(s^{(m)})}{p(s_m)}. \end{aligned} \quad (3.17)$$

D'autre part, d'après la définition de h_- , il existe une constante $c > 0$ telle que

$$\forall k \in \mathbb{N}^* \quad p(s^{(k)}) \leq c\beta^k, \quad \beta \stackrel{\text{def}}{=} \exp(-(1 - \varepsilon^2)h_-). \quad (3.18)$$

Ainsi, en isolant le premier terme de (3.17), on obtient

$$t^r p(s^{(r)}) \delta_r(t^{-1}) \leq 1 + \rho \sum_{m=2}^r \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} \beta^m, \quad (3.19)$$

où $\rho > c$ correspond à l'inverse de la plus petite probabilité invariante. On note

$$q_k(s) = \rho \max_{2 \leq r \leq k} \sum_{m=2}^r \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} \beta^m.$$

La quantité $q_k(s)$ peut être encadrée brutalement par deux constantes :

$$0 \leq q_k(s) \leq \frac{\rho}{1 - \beta}. \quad (3.20)$$

De fait, on choisit de prendre le max sur l'ensemble $\{2 \leq r \leq k\}$ dans la définition de $q_k(s)$ afin d'obtenir une quantité indépendante de r . Rappelons que l'on cherche à déterminer le comportement asymptotique du produit des fonctions génératrices des $(Z_r(s))$ pour $1 \leq r \leq k$ dans l'équation (3.14). Pour simplifier le calcul, on préfère prendre une quantité indépendante de l'indice de sommation. En effet, heuristiquement, le comportement du produit dépend très fortement de la structure de s . Pour k fixé, il est difficile d'apprécier le comportement de toutes les sommes

$$\sum_{m=2}^r \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} \beta^m \quad \text{où} \quad 1 \leq r \leq k.$$

En choisissant de passer au max, on évite de prendre en compte la structure recouvrante de tous les préfixes d'un même mot.

On trouve finalement, par (3.15), (3.16) et (3.19) la majoration

$$\mathbb{E}[t^{Z_r(s)}] \leq 1 - \frac{1}{c\beta^r \left((1-t)^{-1} + \theta \right) + 1 + q_k(s)}, \quad (3.21)$$

ce qui entraîne

$$\prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}] \leq \exp \left[\sum_{r=1}^k \log \left(1 - \frac{1}{c\beta^r \left((1-t)^{-1} + \theta \right) + 1 + q_k(s)} \right) \right].$$

Afin d'établir finement le comportement asymptotique de cette série, on l'approxime par une intégrale en utilisant la monotonie de la fonction

$$r \mapsto \log \left(1 - \frac{1}{c\beta^r \left((1-t)^{-1} + \theta \right) + 1 + q_k(s)} \right).$$

En faisant le changement de variable $y = c\beta^x \left((1-t)^{-1} + \theta \right)$, on obtient

$$\prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}] \leq \exp \left[\frac{1}{\log \beta} \int_{c\beta^k \left((1-t)^{-1} + \theta \right)}^{c \left((1-t)^{-1} + \theta \right)} \log \left(1 - \frac{1}{y + 1 + q_k(s)} \right)^{-1} \frac{dy}{y} \right].$$

Cette intégrale est convergente dans un voisinage de $+\infty$, donc il existe une constante C , indépendante de k et s telle que

$$\prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}] \leq C \exp \left[\frac{1}{\log \beta} \int_{c\beta^k \left((1-t)^{-1} + \theta \right)}^{+\infty} \log \left(1 - \frac{1}{y + 1 + q_k(s)} \right)^{-1} \frac{dy}{y} \right]. \quad (3.22)$$

On déduit de l'égalité élémentaire pour $u \leq v$

$$\left(1 - \frac{u}{y + v} \right)^{-1} = \frac{1 + \frac{v}{y}}{1 + \frac{v-u}{y}},$$

qu'il suffit de s'intéresser au comportement de l'intégrale

$$\int_{a_k}^{+\infty} \log(1 + v/y) \frac{dy}{y} = -\text{Li}_2\left(-\frac{v}{a_k}\right),$$

où $a_k \stackrel{\text{def}}{=} c\beta^k \left((1-t)^{-1} + \theta \right) > 0$ et pour $v > 0$, $\frac{d}{dy} \text{Li}_2\left(-\frac{v}{y}\right) = \frac{1}{y} \log(1 + v/y)$. La notation Li_2 désigne le di-logarithme (voir par exemple Abramowitz et Stegun [1])

$$\text{Li}_2(z) = \sum_{k \geq 1} z^k / k^2.$$

Le di-logarithme est une fonction analytique sur le disque unité qui se prolonge au plan complexe excepté sur la droite $[1, +\infty[$. De plus, au voisinage de $-\infty$,

$$\text{Li}_2(x) = -\frac{1}{2} \log^2(-x) - \zeta(2) + O(1/x), \quad (3.23)$$

ce qui entraîne en particulier que

$$\int_{a_k}^{+\infty} \log\left(1 + \frac{v}{y}\right) \frac{dy}{y} = \frac{1}{2} \log^2\left(\frac{v}{a_k}\right) + \zeta(2) + O(a_k) \quad (a_k \rightarrow 0, a_k > 0).$$

D'autre part, la fonction $\text{Li}_2(x) + \frac{1}{2} \log^2(-x)$ est décroissante sur $] -\infty, 0[$ et donc

$$\text{Li}_2(x) \geq -\frac{1}{2} \log^2(-x) - \zeta(2) \quad (x < 0) \quad (3.24)$$

$$\text{Li}_2(x) \leq -\frac{1}{2} \log^2(-x) - \frac{\zeta(2)}{2} \quad (x < -1)$$

en notant que $\text{Li}_2(-1) = -\frac{\zeta(2)}{2}$. On peut donc calculer la valeur de l'intégrale dans l'équation (3.22)

$$\begin{aligned} \int_{a_k}^{+\infty} \log\left(1 - \frac{1}{y+1+q_k(s)}\right)^{-1} \frac{dy}{y} &= \text{Li}_2\left(-\frac{q_k(s)}{a_k}\right) - \text{Li}_2\left(-\frac{1+q_k(s)}{a_k}\right) \\ &\geq \text{Li}_2\left(-\frac{q_k(s)}{a_k}\right) + \frac{1}{2} \log^2(a_k) + \frac{\zeta(2)}{2}. \end{aligned} \quad (3.25)$$

Le comportement de cette intégrale dépend du comportement asymptotique de $q_k(s)$. Heuristiquement, si $s^{(k)}$ n'est « presque » pas auto-recouvrant, $q_k(s)$ est très proche de 0 ; si au contraire $s^{(k)}$ est « fortement » auto-recouvrant, $q_k(s)$ s'approche d'une constante. Il s'agit de comparer l'asymptotique de $q_k(s)$ avec celle de a_k . Rappelons qu'avec le choix de $t = (1 + c/n)^{-1}$ et de $n = \exp(kh_-(1 - \varepsilon))$, on a

$$a_k = c\beta^k((1-t)^{-1} + \theta) \sim \exp(-kh_-(\varepsilon - \varepsilon^2)). \quad (3.26)$$

On constitue alors deux familles de mots. Dans la première $q_k(s)$ ne tend pas assez vite vers 0, c'est-à-dire à une vitesse $\exp[o(k) \log \beta]$. On choisit arbitrairement la vitesse $z_k \stackrel{\text{def}}{=} \exp(-\sqrt{k})$ satisfaisant cette condition. Sur cette famille, on ne pourra pas majorer très finement le produit des génératrices, mais la perte sera compensée par le fait que peu de mots ont un recouvrement si important. Sur l'autre famille, au contraire, le produit des génératrices sera tout petit.

(i) **Cas des mots $s^{(k)}$ tels que $q_k(s) < z_k$.**

On déduit de (3.24) et de (3.25) que

$$\begin{aligned} \int_{a_k}^{+\infty} \log\left(1 - \frac{1}{y+1+q_k(s)}\right)^{-1} \frac{dy}{y} &\geq -\frac{1}{2} \log^2\left(\frac{z_k}{a_k}\right) + \frac{1}{2} \log^2(a_k) - \frac{\zeta(2)}{2} \\ &\geq k^{3/2} h_-(\varepsilon - \varepsilon^2) + \frac{k^2}{2} - \frac{\zeta(2)}{2} \end{aligned}$$

On déduit des définitions de a_k et z_k ainsi que de la majoration (3.22) que

$$\prod_{j=1}^k \mathbb{E}[t^{Z_j(s)}] \leq C \exp \left[-\frac{\varepsilon}{1+\varepsilon} k^{3/2} + O(k) \right].$$

Rappelons que t a été choisi pour que t^{-n} reste borné. De plus, il y a 4^k mots de longueur k , donc, très grossièrement, en prenant la somme sur tous les mots de longueur k tels que $q_k(s) < z_k$, et en majorant leur nombre par 4^k ,

$$\sum_{s^{(k)} \in \mathcal{A}_k \mid q_k(s) < z_k} \mathbb{P} \left(T_k(s) < \exp(kh_-(1-\varepsilon)) \right) \leq 4^k \exp \left[-\frac{\varepsilon}{1+\varepsilon} k^{3/2} + O(k) \right],$$

qui est le terme général d'une série absolument convergente.

(ii) **Cas des mots $s^{(k)}$ tels que $q_k(s) \geq z_k$.**

Pour cette famille, on considère la majoration de $q_k(s)$ par une constante donnée en (3.20) et on déduit alors de (3.15), (3.16) et (3.19) que

$$\mathbb{E}[t^{Z_r(s)}] \leq 1 - \frac{1}{c\beta^r \left((1-t)^{-1} + \theta \right) + 1 + \rho(1-\beta)^{-1}}.$$

De même que précédemment, en prenant le logarithme du produit, puis en approximant par une intégrale et en faisant le même changement de variable $y = c\beta^x \left((1-t)^{-1} + \theta \right)$, on obtient

$$\prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}] \leq \exp \left[\frac{1}{\log \beta} \int_{c\beta^k \left((1-t)^{-1} + \theta \right)}^{c \left((1-t)^{-1} + \theta \right)} \log \left(1 - \frac{1}{y + 1 + \rho(1-\beta)^{-1}} \right)^{-1} \frac{dy}{y} \right].$$

Cette intégrale est convergente dans un voisinage de $+\infty$, donc il existe une constante C' indépendante de k et s telle que

$$\prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}] \leq C' \exp \left[\frac{1}{\log \beta} \int_{a_k}^{+\infty} \log \left(1 - \frac{1}{y + 1 + \rho(1-\beta)^{-1}} \right)^{-1} \frac{dy}{y} \right].$$

Rappelons que a_k est définie par (3.26). Puisque $x \leq \log(1-x)^{-1}$, on a

$$\prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}] \leq C' \exp \left[\frac{1}{\log \beta} \int_{a_k}^{+\infty} \frac{1}{y + 1 + \rho(1-\beta)^{-1}} \frac{dy}{y} \right].$$

En décomposant en éléments simples puis en intégrant, on obtient

$$\begin{aligned} \int_{a_k}^{+\infty} \frac{1}{y + 1 + \rho(1-\beta)^{-1}} \frac{dy}{y} &= \frac{1}{1 + \rho(1-\beta)^{-1}} \left(\log(a_k + 1 + \rho(1-\beta)^{-1}) - \log(a_k) \right) \\ &= -\frac{\log a_k}{1 + \rho(1-\beta)^{-1}} (1 + o(1)) \end{aligned}$$

puisque a_k tend vers 0. Finalement, grâce à (3.26), il vient

$$\prod_{r=1}^k \mathbb{E}[t^{Z_r(s)}] \leq \exp\left(-\frac{\varepsilon}{1+\varepsilon}k + o(k)\right).$$

Il reste à déterminer le nombre de mots $s^{(k)}$ vérifiant $q_k(s) \geq z_k$, c'est-à-dire le cardinal de l'ensemble

$$E_k \stackrel{\text{def}}{=} \left\{s^{(k)} \mid q_k(s) \geq e^{-\sqrt{k}}\right\}.$$

Par définition de $q_k(s)$, on a

$$E_k \subset \left\{s^{(k)} \mid \exists r \leq k : \rho \sum_{m=2}^r \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} \beta^m \geq e^{-\sqrt{k}}\right\}.$$

Pour alléger un peu les notations, on définit pour $r \leq k$

$$S_r(x) \stackrel{\text{def}}{=} \left\{s^{(k)} \mid \sum_{m=2}^r \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} \beta^m < x\right\}.$$

Si $\ell \in \{2, \dots, r\}$, on a évidemment l'inclusion

$$\bigcap_{m=2}^{\ell} \left\{s^{(k)} \mid \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} = 0\right\} \subset S_r\left(\frac{\beta^{\ell+1}}{1-\beta}\right).$$

On note \bar{B} l'ensemble complémentaire de B dans \mathcal{A}^k . On a alors

$$\bar{S}_r\left(\frac{\beta^{\ell+1}}{1-\beta}\right) \subset \bigcup_{m=2}^{\ell} \left\{s^{(k)} \mid \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} = 1\right\}.$$

Puisque $e^{-\sqrt{k}} = \rho \beta^{\ell+1} (1-\beta)^{-1}$ pour $\ell \stackrel{\text{def}}{=} (-\sqrt{k} + \log(\rho^{-1}(1-\beta)))/\log \beta - 1$, en prenant la contraposée de la précédente inclusion on obtient

$$E_k \subset \bigcup_{r=1}^k \bigcup_{m=2}^{[\ell]+1} \left\{s^{(k)} \mid \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} = 1\right\}.$$

Finalement, on a la majoration

$$|E_k| \leq \sum_{r=1}^k \sum_{m=2}^{[\ell]+1} 4^{m-1} = \mathcal{O}\left(k 4^{\sqrt{k}/\log(\beta^{-1})}\right).$$

Il reste à appliquer le lemme de Borel-Cantelli et faire tendre ε vers zéro pour obtenir

$$\limsup_{n \rightarrow \infty} \frac{\mathcal{L}_n}{\log n} \leq \frac{1}{h_-} \quad \text{p.s.}$$

■

3.4 Convergence en probabilité de la profondeur d'insertion

Dans cette section, on étudie le comportement asymptotique de la profondeur d'insertion notée D_n et de la longueur d'un chemin, choisi aléatoirement et uniformément, notée M_n (voir la Sous-section 3.3.1). D_n est définie comme la longueur du chemin partant de la racine et conduisant au nœud où $W(n)$ est inséré. En d'autres termes, D_n est le nombre de lettres nécessaires à parcourir avant de trouver la position de $W(n)$. Le théorème 3.3.4 a des conséquences immédiates sur le comportement asymptotique de D_n . En effet, puisque $D_n = \ell_n$ lorsque $\ell_{n+1} > \ell_n$, ce qui se produit infiniment souvent presque sûrement puisque $\lim \ell_n = \infty$ p.s., on en déduit que

$$\liminf_{n \rightarrow \infty} \frac{D_n}{\log n} = \liminf_{n \rightarrow \infty} \frac{\ell_n}{\log n} = \frac{1}{h_+}.$$

De même, puisque $D_n = \mathcal{L}_n$ lorsque $\mathcal{L}_{n+1} > \mathcal{L}_n$, on a

$$\limsup_{n \rightarrow \infty} \frac{D_n}{\log n} = \limsup_{n \rightarrow \infty} \frac{\mathcal{L}_n}{\log n} = \frac{1}{h_-}.$$

Théorème 3.4.1.

$$\frac{D_n}{\log n} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \frac{1}{h} \quad \text{et} \quad \frac{M_n}{\log n} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \frac{1}{h}.$$

Remarque 3.4.2. Pour une séquence i.i.d. $U = U_1 U_2 \dots$, dans le cas où les variables aléatoires U_n ne sont pas uniformément distribuées sur l'alphabet $\{A, C, G, T\}$, le théorème 3.4.1 implique que $\frac{D_n}{\log n}$ ne converge pas presque sûrement. En effet, on a clairement

$$\limsup_{n \rightarrow \infty} \frac{D_n}{\log n} = \frac{1}{h_-} > \frac{1}{h} > \frac{1}{h_+} = \liminf_{n \rightarrow \infty} \frac{D_n}{\log n}.$$

Preuve du théorème 3.4.1 Il suffit d'étudier la convergence de D_n puisque, par définition de M_n ,

$$\mathbb{P}(M_n = r) = \frac{1}{n} \sum_{\nu=1}^n \mathbb{P}(D_\nu = r).$$

Soit $\varepsilon > 0$. On prouve le théorème 3.4.1 en établissant la convergence $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0$ où

$$A_n \stackrel{\text{def}}{=} \left\{ U \in \mathcal{A}^{\mathbb{N}} : \left| \frac{D_n}{\log n} - \frac{1}{h} \right| \geq \frac{\varepsilon}{h} \right\}.$$

On a bien évidemment la décomposition

$$\mathbb{P}(A_n) = \mathbb{P}\left(\frac{D_n}{\log n} \geq \frac{1+\varepsilon}{h}\right) + \mathbb{P}\left(\frac{D_n}{\log n} \leq \frac{1-\varepsilon}{h}\right).$$

La définition (3.4) de X_n conduit à l'expression suivante de la profondeur d'insertion

$$D_n = X_{n-1}(W(n)) + 1.$$

Grâce à la dualité entre $X_n(s)$ et $T_k(s)$ explicitée par (3.4), on obtient alors

$$\mathbb{P}\left(\frac{D_n}{\log n} \geq \frac{1+\varepsilon}{h}\right) \leq \mathbb{P}\left(X_{n-1}(W(n)) \geq k-1\right) \leq \mathbb{P}\left(T_{k-1}(W(n)) \leq n-1\right) \quad (3.27)$$

avec $k \stackrel{\text{def}}{=} \lceil \log n(1+\varepsilon)/h \rceil$. On décompose alors le majorant de (3.27) de la façon suivante

$$\mathbb{P}\left(T_{k-1}(W(n)) \leq n-1\right) \leq \mathbb{P}\left(\{T_{k-1}(W(n)) \leq n-1\} \cap B_{k,k_0}\right) + \mathbb{P}(B_{k,k_0}^c),$$

où B_{k,k_0} est défini, pour $k_0 \leq k$, par

$$B_{k,k_0} \stackrel{\text{def}}{=} \bigcap_{k_0 \leq j \leq k} \left\{ U \in \mathcal{A}^{\mathbb{N}} \text{ tel que } \left| \frac{1}{j} \log\left(\frac{1}{p(W(n)^{(j)})}\right) - h \right| \leq \varepsilon^2 h \right\}.$$

Comme la séquence U est stationnaire on a $\mathbb{P}(W(n)^{(k)}) = \mathbb{P}(U^{(k)})$. Le théorème ergodique entraîne alors

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log\left(\frac{1}{p(W(n)^{(k)})}\right) = h \quad \text{p.s.}$$

et ainsi, pour $\eta > 0$ et k_0 assez grand, on a $\mathbb{P}(B_{k,k_0}) \geq 1 - \eta$. On définit également

$$\mathcal{S}_{k,k_0} \stackrel{\text{def}}{=} \bigcap_{k_0 \leq j \leq k} \left\{ s^{(k)} \in \mathcal{A}^k \text{ tels que } \left| \frac{1}{j} \log\left(\frac{1}{p(s^{(j)})}\right) - h \right| \leq \varepsilon^2 h \right\}.$$

Il vient alors pour k_0 et k assez grands

$$\begin{aligned} \mathbb{P}\left(T_{k-1}(W(n)) \leq n-1\right) &\leq \sum_{s^{(k)} \in \mathcal{S}_{k,k_0}} \mathbb{P}\left(W(n)^{(k)} = s^{(k)}, T_{k-1}(s) \leq n-1\right) + \eta \\ &\leq \sum_{s^{(k)} \in \mathcal{S}_{k,k_0}} \mathbb{P}\left(T_{k-1}(s) \leq n-1\right) + \eta \end{aligned}$$

Cette dernière probabilité a déjà été majorée dans la preuve de la Section 3.3.4. On obtient ainsi avec des arguments identiques

$$\sum_{s^{(k)} \in \mathcal{S}_{k,k_0}} \mathbb{P}\left(T_{k-1}(s) \leq n-1\right) = \mathcal{O}\left(k \exp(-\varepsilon(1+\varepsilon)^{-1}k + \log 4 / ((1-\varepsilon^2)h)\sqrt{k})\right). \quad (3.28)$$

On déduit immédiatement de (3.27) et (3.28) que $\mathbb{P}\left(\frac{D_n}{\log n} \geq \frac{1+\varepsilon}{h}\right)$ tend vers zéro quand n tend vers l'infini.

Avec un raisonnement analogue, on montre que $\mathbb{P}\left(\frac{D_n}{\log n} \leq \frac{1-\varepsilon}{h}\right)$ tend vers zéro. En effet, de la majoration

$$\mathbb{P}\left(\frac{D_n}{\log n} \leq \frac{1-\varepsilon}{h}\right) \leq \mathbb{P}\left(X_{n-1}(W(n)) \leq k-1\right) = \mathbb{P}\left(T_k(W(n)) \geq n\right)$$

avec $k \stackrel{\text{def}}{=} \lfloor \log n(1 - \varepsilon)/h \rfloor$, on déduit

$$\mathbb{P}\left(\frac{D_n}{\log n} \leq \frac{1 - \varepsilon}{h}\right) \leq \mathbb{P}\left(\{T_k(W(n)) \geq n\} \cap B_{k,k_0}\right) + \mathbb{P}(B_{k,k_0}^c).$$

Pour tout $\eta > 0$, $\mathbb{P}(B_{k,k_0}^c) \leq \eta$ dès que k_0 est assez grand, et ainsi

$$\begin{aligned} \mathbb{P}\left(T_k(W(n)) \geq n\right) &\leq \sum_{s^{(k)} \in \mathcal{S}_{k,k_0}} \mathbb{P}\left(W(n)^{(k)} = s^{(k)}, T_k(s) \geq n\right) + \eta \\ &\leq \sum_{s^{(k)} \in \mathcal{S}_{k,k_0}} \mathbb{P}\left(T_k(s) \geq n\right) + \eta. \end{aligned}$$

Avec des arguments identiques à ceux de la preuve de la Section 3.3.3, on écrit

$$\sum_{s^{(k)} \in \mathcal{S}_{k,k_0}} \mathbb{P}\left(T_k(s) \leq n\right) = \mathcal{O}\left(4^k \exp(-\kappa n^\theta/2)\right). \quad (3.29)$$

Finalement, on déduit de (3.28) et (3.29) que

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0,$$

ce qui conclut la preuve du théorème 3.4.1. ■

3.5 Expérimentations numériques

Dans une première série d'expériences, on s'intéresse à l'aspect dynamique de la construction de l'arbre. On calcule et représente sur la Figure 3.5 les longueurs $\ell_n/\log n$ et $\mathcal{L}_n/\log n$ ainsi que la profondeur d'insertion $D_n/\log n$ en fonction de n . Sur la courbe du haut, la séquence U est i.i.d. et équiprobable. Pour la courbe du bas, on choisit arbitrairement une loi dans le cas non équiprobable $(p_A, p_C, p_G, p_T) = (0.4, 0.3, 0.2, 0.1)$. Sur les deux graphiques, on voit clairement que la profondeur d'insertion $D_n/\log n$ oscille entre les deux limites $1/h_-$ et $1/h_+$. La limite supérieure est plus souvent dépassée que la limite inférieure, et la convergence de $\mathcal{L}_n/\log n$ est moins rapide que celle de $\ell_n/\log n$. Cependant, on peut noter que la convergence du rapport $\ell_n/\log n$ est très rapide. Ce résultat empirique n'est pas surprenant, il confirme la Remarque 3.3.9. Dans la preuve concernant la convergence des branches les plus courtes, on a pu conserver des majorations assez grossières par rapport à celles qui ont servi à établir le comportement asymptotique des branches les plus longues.

Dans une deuxième série d'expériences, on s'intéresse à l'aspect plus statistique de la convergence. On génère 2 000 séquences i.i.d. de loi $(0.6, 0.1, 0.1, 0.2)$ et de longueur $n = 100\,000$. On représente sur la Figure 3.6 l'histogramme des répartitions de D_n , ℓ_n et \mathcal{L}_n que l'on compare respectivement aux valeurs théoriques $\log n/h$, $\log n/h_+$ et $\log n/h_-$. On retrouve la convergence plus rapide de $\ell_n/\log n$. On note également que les hauteurs sont légèrement plus grandes que leurs valeurs asymptotiques, et que D_n est répartie de manière symétrique autour de $\log n/h$.

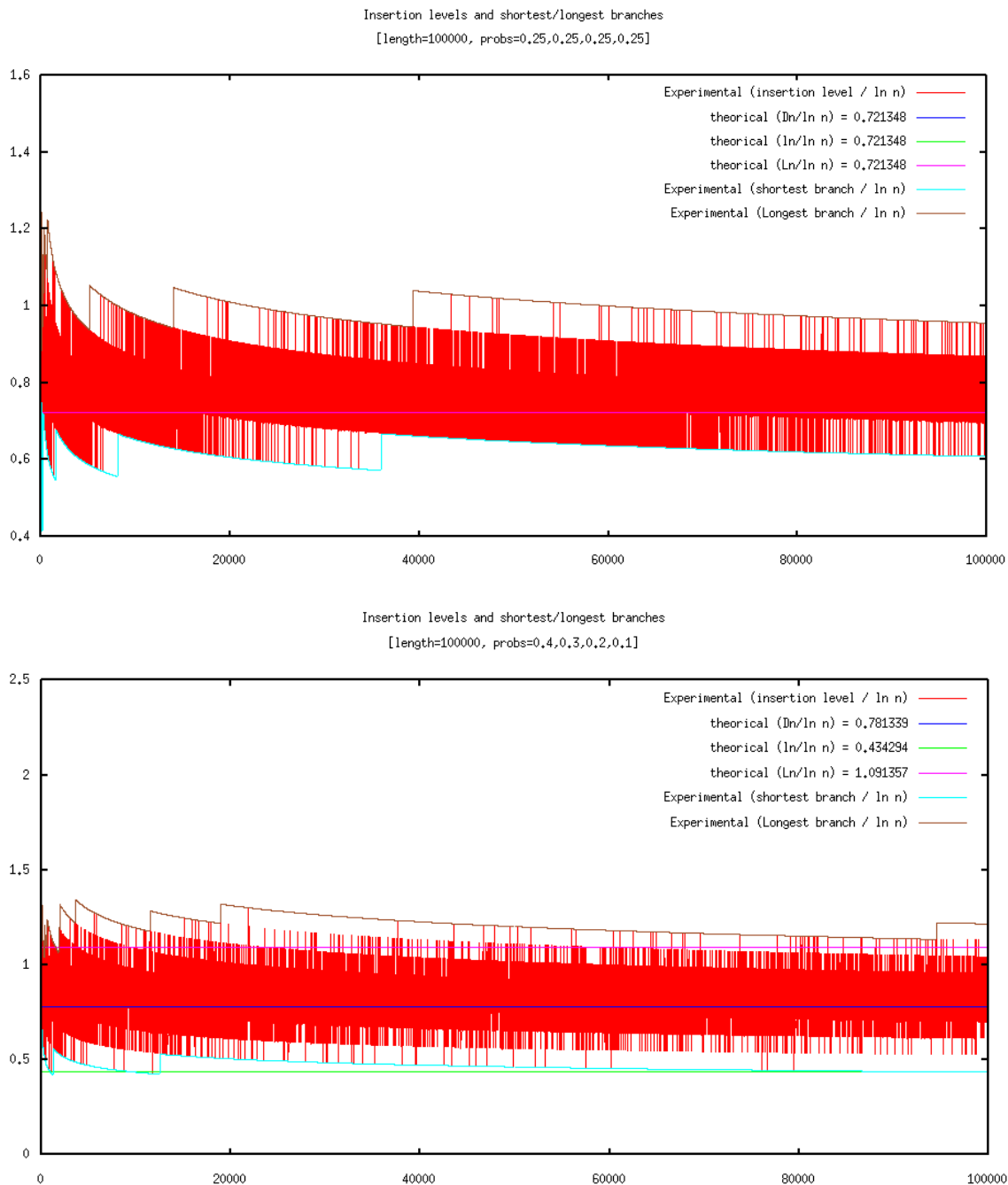


FIG. 3.5: Convergence dynamique des longueurs de branches et de la profondeur d'insertion en fonction de n . Pour une séquence i.i.d. équiprobable (en haut) ou non équiprobable (en bas), on représente les rapports $D_n/\log n$, $\ell_n/\log n$ et $\mathcal{L}_n/\log n$ en fonction de n . On a ajouté les limites $1/h$, $1/h_+$ et $1/h_-$. Dans le cas équiprobable, ces 3 constantes sont égales.

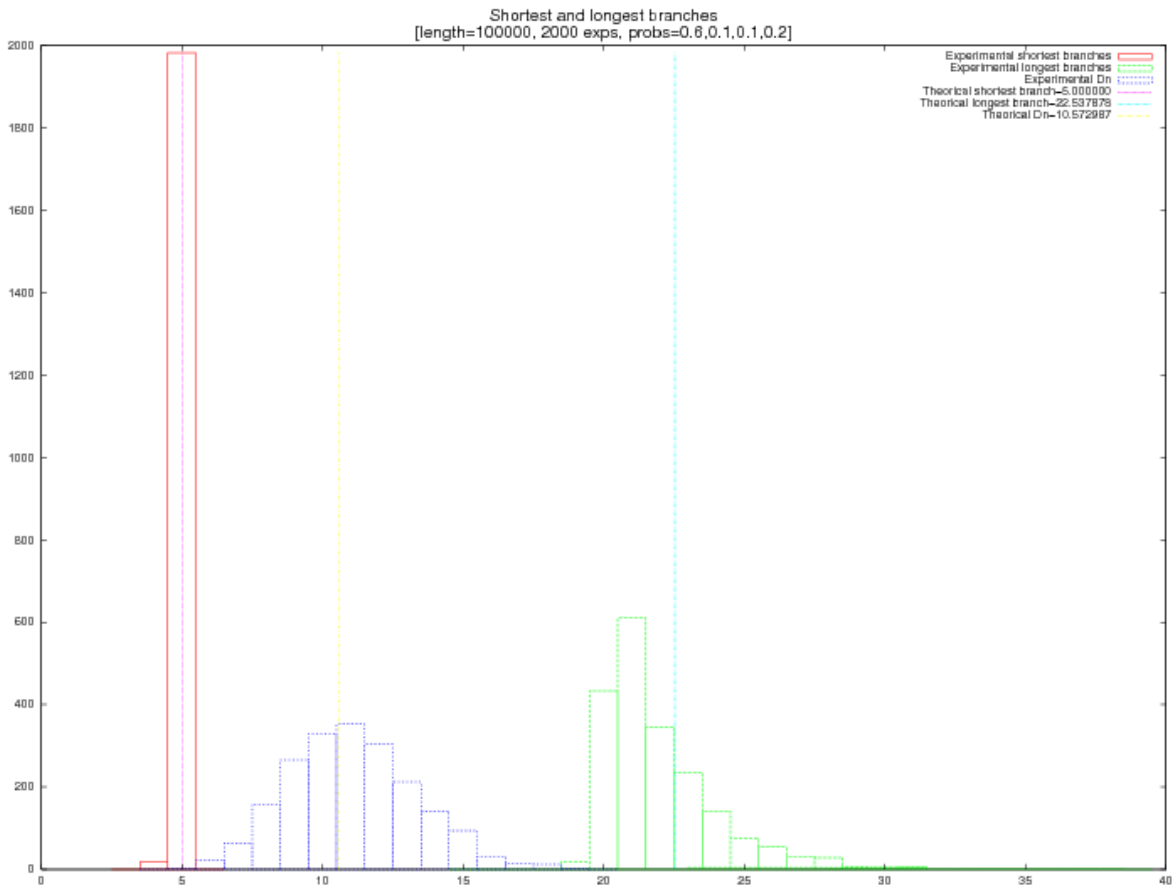


FIG. 3.6: Histogramme des hauteurs \mathcal{L}_n , des longueurs de plus courtes branches ℓ_n et des profondeurs d'insertion D_n , pour $n = 100\,000$ et pour des arbres construits à partir de séquences i.i.d. de loi $(0.6, 0.1, 0.1, 0.2)$. Les valeurs asymptotiques théoriques ont été ajoutées.

3.6 Logiciels développés

La suite de programmes MyCGR indiquée dans la Section 2.4 permet également la manipulation des arbres-CGR définis dans ce chapitre.

Le programme `mycgr_tree.x` permet

- de construire l'arbre-CGR à partir d'une séquence et l'afficher au format Graphviz (Ellson et al. [31]),
- de générer les points correspondants dans le carré,
- d'afficher des informations sur l'arbre, comme le taux de remplissage par niveau, la hauteur, les branches les plus courtes et les plus longues.

L'application graphique `mycgr.x` offre la possibilité de construire des arbres à partir de séquences générées d'après une loi donnée. Les caractéristiques de ces arbres (longueurs des branches les plus courtes et longues, dernier niveau d'insertion, longueur et loi de la séquence d'origine) sont stockées dans une base de données. On peut ensuite afficher le comportement statistique de ces arbres d'une part, et le comportement dynamique de la construction, c'est-à-dire l'évolution des longueurs des branches et du niveau d'insertion en fonction du temps, d'autre part. Les valeurs « attendues » d'après les résultats de la Section 3.5 sont également affichées.

La Figure 3.7 montre une capture d'écran de `mycgr.x` pour la partie concernant les arbres-CGR.

3.7 Domaine de définition de la fonction génératrice de la variable représentant la première occurrence d'un mot

3.7.1 Preuve de l'assertion ii) de la Proposition 3.3.3

Il existe une fonction $K(s_1, s_r, m)$ uniformément bornée par une constante

$$K \stackrel{\text{def}}{=} \sup_{s_1, s_r, m} |K(s_1, s_r, m)|$$

et telle que

$$Q^m(s_1, s_r) - p(s_r) = K(s_1, s_r, m)\gamma^m, \quad (3.30)$$

où γ est la seconde valeur propre de la matrice de transition. Par conséquent,

$$\begin{aligned} |\gamma_r(t) - 1| &= \left| \frac{1-t}{tp(s_r)} \sum_{m \geq 1} K(s_1, s_r, m)(\gamma t)^m \right| \\ &\leq \frac{\gamma K}{\min_u p(u)} \frac{|1-t|}{1-|\gamma t|}. \end{aligned}$$

Le résultat s'obtient en posant $\kappa' \stackrel{\text{def}}{=} |\gamma|K/\min_u p(u)$.

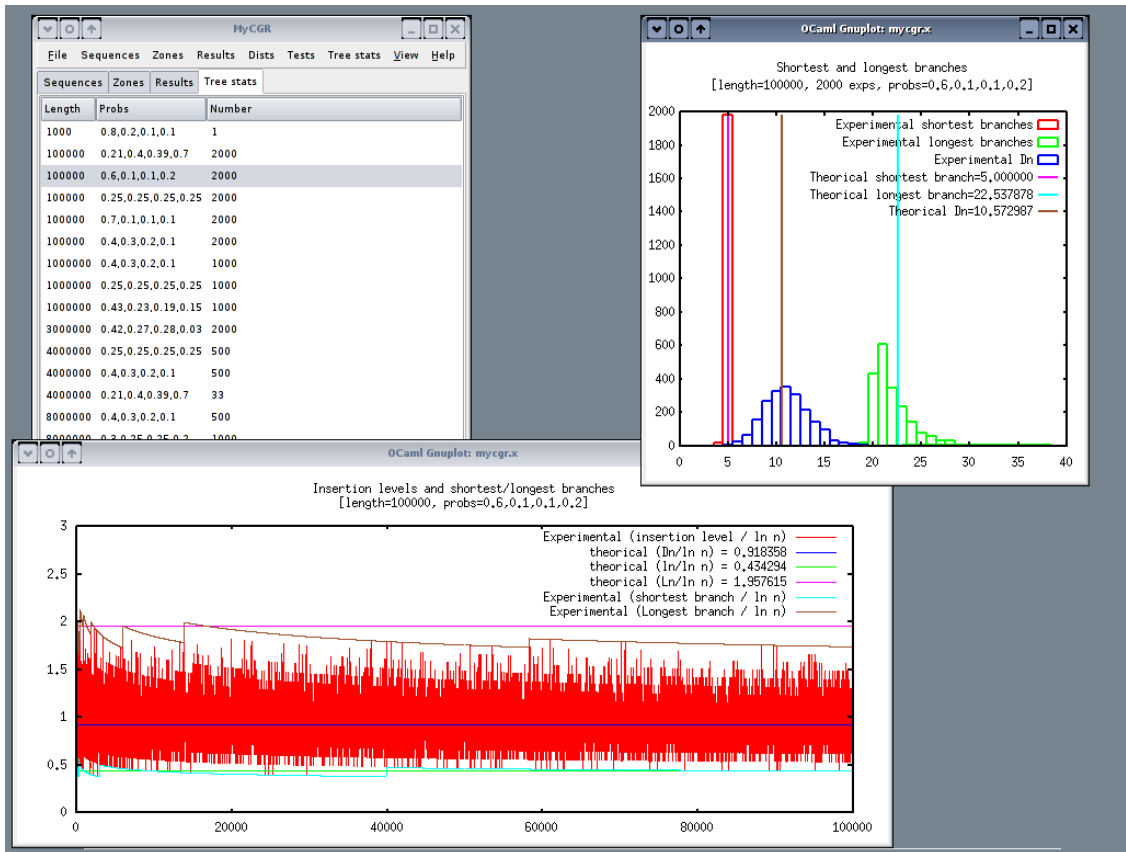


FIG. 3.7: Capture d'écran du programme `mycgr.x` pour la partie arbres-CGR. On distingue la fenêtre principale, à gauche, montrant les nombres d'arbres construits selon des probabilités et des longueurs de séquences données. En sélectionnant l'un de ces groupes d'arbres, on peut afficher la fenêtre de droite qui montre le comportement statistique de ces arbres, ainsi que les valeurs asymptotiques d'après les résultats de la Section 3.5. La fenêtre du bas affiche le comportement dynamique de la construction d'un arbre construit à partir d'une séquence de loi et de longueur données.

3.7.2 Preuve de l'assertion i)

Sur le disque unité $|t| < 1$, la somme

$$S(t) \stackrel{\text{def}}{=} \frac{1}{t} \sum_{m \geq 1} Q^m(s_1, s_r) t^m \quad (3.31)$$

est bien définie et on a la décomposition

$$\frac{1-t}{p(s_r)t} \sum_{m \geq 1} Q^m(s_1, s_r) t^m = 1 + \frac{1-t}{p(s_r)t} \sum_{m \geq 1} [Q^m(s_1, s_r) - p(s_r)] t^m.$$

La fonction

$$\sum_{m \geq 1} [Q^m(s_1, s_r) - p(s_r)] t^m$$

admet un prolongement analytique dans la région $|\gamma t| < 1$. Donc la série

$$\frac{1-t}{tp(s_r)} \sum_{m \geq 1} Q^m(s_1, s_r) t^m$$

est convergente sur le même domaine. $S'(t)$ est analytique sur le disque $|\gamma t| < 1$. Il reste à déterminer les zéros de

$$\begin{aligned} D(t) \stackrel{\text{def}}{=} p(s^{(r)}) t^r &+ \frac{(1-t)p(s^{(r)}) t^r}{p(s_r)t} \sum_{z \geq 1} t^z [Q^z(s_1, s_r) - p(s_r)] \\ &+ (1-t) \left[1 + \sum_{j=1}^{r-1} t^j p(s^{(j)}) \mathbb{1}_{\{s_{r-j} \dots s_1 = s_r \dots s_{j+1}\}} \right]. \end{aligned}$$

On suppose que t est une racine réelle de $D(t)$, avec $0 < t < 1$, alors on a

$$\begin{aligned} 0 &< \frac{(1-t)p(s^{(r)}) t^r}{p(s_r)t} \sum_{z \geq 1} t^z Q^z(s_1, s_r) \\ &= (t-1) \left[1 + \sum_{j=1}^{r-1} t^j p(s^{(j)}) \mathbb{1}_{\{s_{r-j} \dots s_1 = s_r \dots s_{j+1}\}} \right] < 0. \end{aligned}$$

Par conséquent, il est évident qu'il n'y a pas de racine de $D(t)$ dans $]0, 1[$. De plus, on vérifie facilement que 0 et 1 ne sont pas des racines de $D(t)$. On s'intéresse maintenant aux racines de la forme $t = 1 + \varepsilon$ avec $\varepsilon > 0$, c'est-à-dire telles que

$$\varepsilon = \frac{(1+\varepsilon)^r p(s^{(r)}) \left(1 - \frac{\varepsilon}{p(s_r)(1+\varepsilon)} \sum_{z \geq 1} t^z [Q^z(s_1, s_r) - p(s_r)] \right)}{\left(1 + \sum_{j=1}^{r-1} (1+\varepsilon)^j p(s^{(j)}) \mathbb{1}_{\{s_{r-j} \dots s_1 = s_r \dots s_{j+1}\}} \right)}.$$

On a alors l'inégalité

$$\begin{aligned}\varepsilon &\geq \frac{cp(s^{(r)})}{1 + \sum_{k=1}^{r-1} p(s^{(r-k)})} \\ &\geq \kappa p(s^{(r)}).\end{aligned}$$

Finalement, la fonction génératrice de $Y_r(s)$ notée $\Phi(s^{(r)}, t)$ est au moins définie sur $[0, 1 + \kappa p(s^{(r)})[$.

Chapitre 4

Convergence des moments dans le Théorème de la limite centrale pour les martingales vectorielles

L'objectif de ce chapitre est d'établir de nouvelles propriétés de convergence presque sûre de transformées de martingales vectorielles. On montre en particulier que, sous certaines conditions de régularité du processus croissant et sous certaines conditions de moments sur la martingale, il y a convergence des moments normalisés de tout ordre dans le théorème de la limite centrale presque sûr (TLCPS) pour les martingales vectorielles. Les résultats de convergence sont appliqués aux modèles de régression linéaire, ainsi qu'aux processus de branchement, afin d'établir de nouvelles propriétés asymptotiques sur les erreurs d'estimation et de prédiction.

Sommaire

4.1	Théorème de la limite centrale presque sûr et état de l'art	102
4.1.1	Cas scalaire	102
4.1.2	Cas vectoriel	103
4.1.3	Applications	105
4.2	Convergence des moments dans le TLCPS pour les martingales vectorielles	106
4.3	Applications statistiques	108
4.3.1	Estimation des moments, erreurs d'estimation et de prédiction .	109
4.3.2	Modèles autorégressifs linéaires	110
4.3.3	Processus de branchement avec immigration	111
4.3.4	Un lien entre la CGR et les processus RCA	114
4.3.5	Sur l'estimateur des moindres carrés pondérés	115
4.4	Preuves	118
4.4.1	Preuve du théorème 4.2.1	118
4.4.2	Preuve du corollaire 4.2.4	131
4.4.3	Preuve du théorème 4.2.5	131
4.4.4	Preuve du corollaire 4.3.1	132
4.4.5	Preuve du corollaire 4.3.2	133

4.1 Théorème de la limite centrale presque sûr et état de l'art

4.1.1 Cas scalaire

Soit (ξ_n) une suite de variables indépendantes et de même loi, avec $\mathbb{E}[\xi_n] = 0$ et $\mathbb{E}[\xi_n^2] = \sigma^2$. Définissons $Z_n \stackrel{\text{def}}{=} \xi_1 + \dots + \xi_n$. Alors, d'après le théorème de la limite centrale, pour toute fonction h continue bornée,

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[h \left(\frac{Z_n}{\sqrt{n}} \right) \right] = \int_{\mathbb{R}} h(x) dG(x),$$

où G est la mesure gaussienne $\mathcal{N}(0, \sigma^2)$. On a également, par le TLCPS, que

$$\frac{1}{\log n} \sum_{k=1}^n \frac{1}{k} \delta_{\frac{Z_k}{\sqrt{k}}} \Longrightarrow G \quad \text{p.s.}$$

En d'autres termes, pour toute fonction h continue bornée,

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n \frac{1}{k} h \left(\frac{Z_k}{\sqrt{k}} \right) = \int_{\mathbb{R}} h(x) dG(x) \quad \text{p.s.}$$

Ce théorème a été démontré par Brosamler [12] et Schatte [84; 85], et dans sa forme présente par Lacey [57]. Le théorème de la limite centrale presque sûr a aussi été établi dans un cadre martingales par Chaâbane [18; 19], Chaâbane et Maâouia [20] et Lifshits [63; 64]. Plus précisément, le contexte martingales est défini de la manière suivante. Soit (ε_n) une suite de différences de martingale adaptée à une filtration $\mathbb{F} \stackrel{\text{def}}{=} (\mathcal{F}_n)$. Soit (Φ_n) une suite de variables aléatoires adaptée à \mathbb{F} . La transformée de martingale réelle (M_n) est définie, pour tout $n \geq 1$, par

$$M_n \stackrel{\text{def}}{=} \sum_{k=1}^n \Phi_{k-1} \varepsilon_k.$$

On définit également le coefficient d'explosion f_n associé à Φ_n par

$$f_n \stackrel{\text{def}}{=} \frac{\Phi_n^2}{s_n} \quad \text{avec} \quad s_n \stackrel{\text{def}}{=} \sum_{k=0}^n \Phi_k^2.$$

Théorème 4.1.1. *On suppose que $\mathbb{E}[\varepsilon_{n+1}^2 \mid \mathcal{F}_n] = \sigma^2$ p.s. et on note (U_n) une suite positive prévisible (i.e. \mathcal{F}_{n-1} -mesurable) telle que*

$$\begin{aligned} & \lim_{n \rightarrow \infty} U_n^{-2} s_{n-1} = 1 \quad \text{p.s.} \\ \forall \delta > 0 & \quad \sum_{n=1}^{\infty} U_n^{-2} \mathbb{E}[(M_n - M_{n-1})^2 \mathbf{1}_{\{|M_n - M_{n-1}| > \delta U_n\}} \mid \mathcal{F}_{n-1}] < \infty \quad \text{p.s.} \\ \exists a > 0 & \quad \sum_{n=1}^{\infty} U_n^{-2a} \mathbb{E}[|M_n - M_{n-1}|^{2a} \mathbf{1}_{\{|M_n - M_{n-1}| \leq U_n\}} \mid \mathcal{F}_{n-1}] < \infty \quad \text{p.s.} \end{aligned}$$

Alors, (M_n) satisfait le TLCPs

$$\frac{1}{\log U_{n+1}^2} \sum_{k=1}^n \left(\frac{U_{k+1}^2 - U_k^2}{U_{k+1}^2} \right) \delta_{M_k/U_k} \implies G \quad p.s.$$

On peut facilement vérifier que, sous les hypothèses du théorème 4.1.1, U_{n+1}^2 est presque sûrement équivalent à U_n^2 et donc le coefficient d'explosion f_n tend vers zéro p.s. De plus, en choisissant $U_n^2 = s_{n-1}$, on obtient alors que, pour toute fonction h continue bornée,

$$\lim_{n \rightarrow \infty} \frac{1}{\log s_n} \sum_{k=1}^n f_k h\left(\frac{M_k}{\sqrt{s_{k-1}}}\right) = \int_{\mathbb{R}} h(x) dG(x) \quad p.s. \quad (4.1)$$

Le théorème 4.1.1 est une version simplifiée du TLCPs pour les martingales de Chaâbane [18]. Une question naturelle est de se demander si ce théorème reste vrai pour des fonctions h non bornées. Bercu [8] démontre que, sous une hypothèse de moment conditionnel d'ordre strictement plus grand que $2p$ sur (ε_n) , la convergence (4.1) est vérifiée pour toute fonction h telle que $h(x) = x^{2p}$ avec $p \geq 1$.

Théorème 4.1.2 (Convergence des moments dans le TLCPs scalaire). *On suppose que $\mathbb{E}[\varepsilon_{n+1}^2 | \mathcal{F}_n] = \sigma^2$ p.s. et que f_n tend vers zéro p.s. S'il existe un entier $p \geq 1$ et un réel $a > 2p$ tels que*

$$\sup_{n \geq 0} \mathbb{E}[\varepsilon_{n+1}^a | \mathcal{F}_n] < \infty \quad p.s.,$$

alors on a

$$\lim_{n \rightarrow \infty} \frac{1}{\log s_n} \sum_{k=1}^n f_k \left(\frac{M_k^2}{s_{k-1}} \right)^p = \frac{\sigma^{2p} (2p)!}{2^p p!} \quad p.s. \quad (4.2)$$

Remarque 4.1.3. On reconnaît le moment d'ordre $2p$ de la loi gaussienne $\mathcal{N}(0, \sigma^2)$ dans la limite donnée en (4.2).

4.1.2 Cas vectoriel

Que peut-on dire dans le cas vectoriel? Dans ce chapitre, nous établissons un équivalent du théorème 4.1.2 de Bercu [8] dans un contexte multidimensionnel.

Soit (M_n) une martingale à valeurs dans \mathbb{R}^d , adaptée à une filtration \mathbb{F} . On note $(\langle M \rangle_n)$ son processus croissant. Le théorème ci-dessous est une version du théorème de la limite centrale pour les martingales discrètes (voir par exemple Hall et Heyde [46]).

Théorème 4.1.4 (Théorème de la limite centrale pour martingales). *On suppose que (M_n) est une martingale de carré intégrable. De plus, on suppose qu'il existe une*

suite déterministe croissante (α_n) qui tend vers l'infini et telle que

- i) $\alpha_n^{-1} \langle M \rangle_n \xrightarrow[n \rightarrow \infty]{P} \Gamma,$
- ii) La condition de Lindeberg est satisfaite, i.e. pour tout $\varepsilon > 0,$

$$\alpha_n^{-1} \sum_{k=1}^n \mathbb{E} \left[\| M_k - M_{k-1} \|^2 \mathbb{1}_{\{\| M_k - M_{k-1} \| \geq \varepsilon \alpha_n^{1/2}\}} | \mathcal{F}_{k-1} \right] \xrightarrow[n \rightarrow \infty]{P} 0.$$

Alors on a

$$\alpha_n^{-1} M_n \xrightarrow[n \rightarrow \infty]{p.s.} 0 \quad \text{et} \quad \alpha_n^{-1/2} M_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma).$$

De plus, si la matrice Γ est inversible, on a

$$\sqrt{\alpha_n} \langle M \rangle_n^{-1} M_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma^{-1}).$$

Remarque 4.1.5. La condition de Lindeberg est vérifiée si les accroissements de la martingale sont bornés. Plus généralement, elle est vérifiée si la condition de Lyapunov

$$\alpha_n^{-1} \sum_{k=1}^n \mathbb{E} \left[\| M_k - M_{k-1} \|^a | \mathcal{F}_{k-1} \right] \xrightarrow[n \rightarrow \infty]{P} 0$$

est satisfaite, avec $a > 2.$

Une approche du TLCPS pour les martingales vectorielles discrètes a été développée dans Chaàbane et al. [21]. Le théorème 4.1.6 en donne une version simplifiée.

Théorème 4.1.6. On suppose que (M_n) est une martingale de carré intégrable et qu'il existe une suite **déterministe** (U_n) de matrices réelles inversibles satisfaisant les conditions de croissance régulière suivantes : pour tout entier $n,$ la matrice U_n est de la forme $U_n = \alpha_n I_d$ où α_n est une suite croissante vers l'infini avec $\alpha_n \sim \alpha_{n-1}.$ Sous les hypothèses

- i) $U_n^{-1} \langle M \rangle_n U_n^{-1} \xrightarrow[n \rightarrow \infty]{p.s.} C,$ où C est une matrice aléatoire ou non,
- ii) $\forall \delta > 0,$
$$\sum_{k=1}^n \mathbb{E} \left[\| U_n^{-1} (M_k - M_{k-1}) \|^2 \mathbb{1}_{\{\| U_n^{-1} (M_k - M_{k-1}) \| \geq \delta\}} | \mathcal{F}_{k-1} \right] \xrightarrow[n \rightarrow \infty]{p.s.} 0,$$

(M_n) satisfait le TLCPS

$$\frac{1}{\log(\det U_n)^2} \sum_{k=1}^n \left(1 - \left(\frac{\det U_k}{\det U_{k+1}} \right)^2 \right) \delta_{U_k^{-1} M_k} \implies Y \quad p.s.,$$

où Y est la loi mélange $C^{1/2}G,$ G étant un vecteur gaussien standard indépendant de $C.$ Si C n'est pas aléatoire, la mesure limite est la loi gaussienne $\mathcal{N}(0, C).$

On montre dans la suite que, sous des hypothèses appropriées proches de celles de Chaâbane et al. [21], mais en remplaçant les poids déterministes U_n par la racine du processus croissant, il y a convergence des moments dans le TLCPS pour les martingales vectorielles. Soit (Φ_n) une suite de vecteurs aléatoires de \mathbb{R}^d , adaptée à la filtration \mathbb{F} . On définit la transformée de martingales vectorielle

$$M_n \stackrel{\text{def}}{=} M_0 + \sum_{k=1}^n \Phi_{k-1} \varepsilon_k,$$

où M_0 peut être arbitrairement choisie. On note également

$$S_n \stackrel{\text{def}}{=} \sum_{k=0}^n \Phi_k \Phi_k^t + S, \quad (4.3)$$

où S est une matrice définie positive, symétrique et déterministe et, pour toute matrice A , A^t désigne la transposée de A . On peut remarquer que, si $\mathbb{E}[\varepsilon_{n+1}^2 \mid \mathcal{F}_n] = \sigma^2$ p.s., le processus croissant de (M_n) est la matrice $\langle M \rangle_n = \sigma^2 S_n$. On définit également le coefficient d'explosion associé à (Φ_n) par

$$f_n \stackrel{\text{def}}{=} \Phi_n^t S_n^{-1} \Phi_n = \frac{d_n - d_{n-1}}{d_n}, \quad (4.4)$$

où $d_n \stackrel{\text{def}}{=} \det(S_n)$.

4.1.3 Applications

À partir de nouvelles propriétés asymptotiques presque sûres pour les puissances de transformées de martingales vectorielles, on établit des résultats de convergence sur les erreurs d'estimation et de prédiction associées aux modèles de régression linéaire. Ils sont définis, pour tout $n \geq 1$, par la relation

$$X_{n+1} = \theta^t \Phi_n + \varepsilon_{n+1}, \quad (4.5)$$

où $\theta \in \mathbb{R}^d$ est le paramètre inconnu du modèle. Les variables X_n , Φ_n , et ε_n sont respectivement l'observation scalaire, le vecteur de régression et le bruit scalaire du système. En particulier, on illustre les résultats sur les modèles autorégressifs linéaires et les processus de branchement avec immigration.

Pour une suite d'estimateurs $(\hat{\theta}_n)$ de θ , on s'intéresse à la performance asymptotique de $\hat{\theta}_n^t \Phi_n$ comme prédicteur de X_{n+1} . Plus précisément, on se concentre sur l'erreur de prédiction $X_{n+1} - \hat{\theta}_n^t \Phi_n$ et sur l'erreur d'estimation $\hat{\theta}_n - \theta$. Il est plus approprié (voir par exemple Goodwin et Sin [41]) de considérer les erreurs cumulées de prédiction et d'estimation, respectivement définies, pour tout $p \geq 1$, par

$$C_n(p) = \sum_{k=0}^{n-1} (X_{k+1} - \hat{\theta}_k^t \Phi_k)^{2p} \quad (4.6)$$

et

$$G_n(p) = \sum_{k=1}^n k^{p-1} \|\widehat{\theta}_k - \theta\|^{2p}. \quad (4.7)$$

Dans le cas scalaire $d = 1$, sous des hypothèses de moment appropriées, des résultats asymptotiques sur $C_n(p)$ et $G_n(p)$ ont été établis par Bercu [8], en utilisant l'estimateur des moindres carrés

$$\widehat{\theta}_n = S_{n-1}^{-1} \sum_{k=1}^n \Phi_{k-1} X_k. \quad (4.8)$$

Notre but est d'étendre ces résultats au cadre multidimensionnel. Dans ce contexte, Duflo [28], Wei [91] prouvent des résultats asymptotiques pour $C_n(p)$ et $G_n(p)$ mais seulement dans le cas $p = 1$. Pour la consistance forte de l'estimateur des moindres carrés, on peut également consulter les travaux de Lai et Wei [58], Wei [90]. Ces résultats sont étendus dans Duflo et al. [29]. D'autre part, le comportement asymptotique de l'estimateur empirique de la covariance associée au modèle (4.5) est étudié dans Duflo et al. [29], Lai et Wei [58; 59], Wei [90].

4.2 Convergence des moments dans le TLCPS pour les martingales vectorielles

Dans cette section, on généralise le théorème 4.1.2 au cadre vectoriel, en montrant que sous des hypothèses de moment appropriées, et sous des hypothèses de convergence raisonnables du processus croissant, il y a convergence des moments normalisés de tout ordre dans le TLCPS pour les martingales vectorielles.

Pour alléger les notations, on définit les trois hypothèses sur le bruit (ε_n) qui sont utilisées dans les théorèmes et corollaires suivants. On note respectivement (H_{2p}) , (H_{2p+}) et (C_{2p}) les assertions : la suite (ε_n) est une différence de martingale telle que

$$(H_{2p}) \quad \sup_{n \geq 0} \mathbb{E}[\varepsilon_{n+1}^{2p} | \mathcal{F}_n] < \infty \quad \text{p.s.} \quad (4.9)$$

$$(H_{2p+}) \quad \sup_{n \geq 0} \mathbb{E}[|\varepsilon_{n+1}|^a | \mathcal{F}_n] < \infty \quad \text{p.s.} \quad \text{pour un réel } a > 2p, \quad (4.10)$$

$$(C_{2p}) \quad \forall n \geq 0, \quad \mathbb{E}[\varepsilon_{n+1}^{2p} | \mathcal{F}_n] \stackrel{\text{def}}{=} \sigma(2p) < \infty \quad \text{p.s.} \quad (4.11)$$

On note $\sigma(2) = \sigma^2$.

Théorème 4.2.1. *On suppose que (ε_n) est une différence de martingale satisfaisant (C_2) et (H_{2p+}) pour un entier $p \geq 1$. De plus, on suppose que le coefficient d'explosion f_n tend vers zéro p.s. et qu'il existe une suite aléatoire positive (α_n) croissante vers l'infini et une matrice inversible L telles que*

$$\lim_{n \rightarrow \infty} \alpha_n^{-1} S_n = L \quad \text{p.s.} \quad (4.12)$$

Alors on a presque sûrement

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n f_k (M_k^t S_{k-1}^{-1} M_k)^p = \ell(p) \stackrel{\text{def}}{=} d \sigma^{2p} \prod_{j=1}^{p-1} (d + 2j). \quad (4.13)$$

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n [(M_k^t S_{k-1}^{-1} M_k)^p - (M_k^t S_k^{-1} M_k)^p] = \lambda(p) \stackrel{\text{def}}{=} \frac{p}{d} \ell(p). \quad (4.14)$$

Démonstration La preuve du théorème 4.2.1 est donnée dans la Section 4.4. ■

Remarque 4.2.2. L'hypothèse (4.12) entraîne la convergence du coefficient d'explosion f_n vers zéro p.s. si et seulement si $\alpha_n \sim \alpha_{n-1}$ p.s.

Remarque 4.2.3. La limite $\ell(p)$ est le moment d'ordre $2p$ de la norme d'un vecteur gaussien $\mathcal{N}(0, \sigma^2 I_d)$. Rappelons brièvement le calcul permettant de l'établir. On définit la transformée de Laplace de cette norme pour tout $t \in \mathbb{R}$ avec $1 - 2\sigma^2 t > 0$, par

$$\Lambda(t) \stackrel{\text{def}}{=} \mathbb{E}[e^{t\|X\|^2}] = \prod_{i=1}^d \mathbb{E}[e^{tX_i^2}] = (1 - 2\sigma^2 t)^{-d/2}.$$

Un petit raisonnement par récurrence permet de montrer que la dérivée $p^{\text{ième}}$ de Λ vérifie

$$\Lambda^{(p)}(t) = \ell(p) (1 - 2\sigma^2 t)^{\frac{d+2p}{2}}.$$

Finalement, on retrouve l'expression des moments

$$\mathbb{E}[\|X\|^{2p}] = \Lambda^{(p)}(0) = \ell(p).$$

Par conséquent, le théorème 4.2.1 établit bien la convergence des moments d'ordre $2p$ de la norme du vecteur $S_{n-1}^{-1/2} M_n$ dans le TLCPS. Par rapport à la version de Chaâbane et al. [21], la normalisation déterministe U_n est remplacée par la normalisation aléatoire dépendant du processus croissant.

En posant $M_0 = -S\theta$, on déduit de (4.5) et de (4.8) que

$$\widehat{\theta}_n - \theta = S_{n-1}^{-1} M_n. \quad (4.15)$$

On définit

$$\pi_n \stackrel{\text{def}}{=} (\widehat{\theta}_n - \theta)^t \Phi_n = X_{n+1} - \widehat{\theta}_n^t \Phi_n - \varepsilon_{n+1}. \quad (4.16)$$

Alors les deux égalités précédentes (4.15) et (4.16) permettent d'écrire

$$\pi_n^2 = M_n^t S_{n-1}^{-1} \Phi_n \Phi_n^t S_{n-1}^{-1} M_n.$$

Par ailleurs, en appliquant la formule de Riccati, on a

$$S_{n-1}^{-1} = S_n^{-1} + (1 - f_n) S_{n-1}^{-1} \Phi_n \Phi_n^t S_{n-1}^{-1}$$

et par conséquent

$$a_n(1) \stackrel{\text{def}}{=} M_n^t S_{n-1}^{-1} M_n - M_n^t S_n^{-1} M_n = (1 - f_n) \pi_n^2. \quad (4.17)$$

Il est souvent difficile d'obtenir des résultats asymptotiques sur le coefficient d'explosion f_n . Dans les modèles considérés dans ce chapitre, f_n converge p.s. vers zéro. Pour étudier le comportement asymptotique de $G_n(p)$ et $C_n(p)$, on va utiliser des propriétés asymptotiques de $a_n(1)^p$, sous des hypothèses de moments appropriées.

Corollaire 4.2.4. *Sous les hypothèses du théorème 4.2.1, on a également*

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n a_k(1)^p = \begin{cases} 0 & \text{si } p > 1, \\ \sigma^2 & \text{si } p = 1. \end{cases} \quad (4.18)$$

Démonstration Voir Section 4.4. ■

Plus généralement, sans faire l'hypothèse de moment d'ordre 2 constant, on peut aussi déterminer la vitesse de convergence de $\sum f_k \|S_{k-1}^{-1/2} M_k\|^{2p}$.

Théorème 4.2.5. *On suppose que (ε_n) est une différence de martingale satisfaisant (H_{2p+}) pour un entier $p \geq 1$. De plus, on suppose que le coefficient d'explosion f_n tend vers zéro p.s. et qu'il existe une suite aléatoire α_n positive croissante vers l'infini et une matrice inversible L vérifiant (4.12). Alors on a presque sûrement*

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n f_k (M_k^t S_{k-1}^{-1} M_k)^p = \mathcal{O}(1). \quad (4.19)$$

Démonstration Voir Section 4.4. ■

Dans les modèles étudiés ici, l'hypothèse de convergence (4.12) est souvent vérifiée.

4.3 Applications statistiques

Une application possible du théorème 4.2.1 concerne le modèle de régression linéaire défini, pour tout $n \geq 1$, par

$$X_{n+1} = \theta^t \Phi_n + \varepsilon_{n+1}.$$

Le but de cette sous-section est d'établir le comportement de $C_n(p)$ et $G_n(p)$ donnés par (4.6) et (4.7). Du même coup, on estime les moments d'ordre $2p$ du bruit (ε_n) .

4.3.1 Estimation des moments, erreurs d'estimation et de prédiction

Si le bruit (ε_n) vérifie (H_{2p+}) et (C_2) , alors on a presque sûrement

$$\Delta_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k=1}^n \varepsilon_k^2 \longrightarrow \sigma^2.$$

Sous les hypothèses du théorème 4.2.1 pour $p = 1$, on déduit alors de (4.18) que l'estimateur de σ^2

$$\Gamma_n = \frac{1}{n} \sum_{k=0}^{n-1} (X_{k+1} - \hat{\theta}_k^t \Phi_k)^2$$

est consistant avec

$$\lim_{n \rightarrow \infty} \frac{n}{\log d_n} (\Gamma_n - \Delta_n) = \sigma^2 \quad \text{p.s.}$$

On peut maintenant proposer des estimateurs consistants des moments d'ordre supérieur de (ε_n) . Pour $p \geq 1$,

$$\Gamma_n(2p) = \frac{1}{n} \sum_{k=0}^{n-1} (X_{k+1} - \hat{\theta}_k^t \Phi_k)^{2p} \quad (4.20)$$

est un estimateur naturel du moment $\sigma(2p)$ d'ordre $2p$ du bruit (ε_n) . On peut remarquer que $n\Gamma_n(2p) = C_n(p)$. Le corollaire suivant donne des résultats asymptotiques pour $\Gamma_n(2p)$.

Corollaire 4.3.1. *On suppose qu'il existe $p \geq 2$ vérifiant (C_{2p}) . De plus, on suppose qu'il existe une suite aléatoire (α_n) positive, croissante vers l'infini, ainsi qu'une matrice inversible L telles que l'hypothèse de convergence (4.12) soit vérifiée. On suppose également que f_n tend vers zéro p.s. Alors pour tout entier q vérifiant $1 \leq q \leq p$ et (C_{2q}) , $\Gamma_n(2q)$ est un estimateur consistant de $\sigma(2q)$ et*

$$\left(\Gamma_n(2q) - \frac{1}{n} \sum_{k=1}^n \varepsilon_k^{2q} \right)^2 = \mathcal{O}\left(\frac{\log d_n}{n}\right) \quad \text{p.s.} \quad (4.21)$$

Démonstration La preuve est donnée en Section 4.4. ■

On peut maintenant déduire du corollaire 4.3.1 le comportement asymptotique de $C_n(q)$. Sous les hypothèses du corollaire 4.3.1, la convergence (4.21) implique que $C_n(q)/n$ converge p.s. vers $\sigma(2q)$. De plus, si (ε_n) a un moment conditionnel fini d'ordre $a > 2q$, on déduit du lemme de Chow (voir par exemple Duflo [28, p. 22]) que pour c vérifiant $2qa^{-1} < c < 1$, on a

$$\left| \frac{1}{n} \sum_{k=1}^n \varepsilon_k^{2q} - \sigma(2q) \right| = o(n^{c-1}) \quad \text{p.s.} \quad (4.22)$$

Par conséquent, dès que $\log d_n = o(n^c)$, les convergences (4.21) et (4.22) impliquent

$$\left| \frac{1}{n} C_n(q) - \sigma(2q) \right|^2 = o(n^{c-1}) \quad \text{p.s.}$$

Avant d'obtenir des résultats sur l'erreur d'estimation cumulée $G_n(p)$, on énonce un autre corollaire du théorème 4.2.1.

Corollaire 4.3.2. *Sous les hypothèses du théorème 4.2.1, on a*

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n f_k \left((\hat{\theta}_k - \theta)^t S_k (\hat{\theta}_k - \theta) \right)^p = \ell(p) \quad \text{p.s.} \quad (4.23)$$

De plus, on suppose qu'il existe une matrice inversible L telle que

$$\lim_{n \rightarrow +\infty} \frac{1}{n} S_n = L \quad \text{p.s.} \quad (4.24)$$

Alors on a aussi

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n k^{p-1} \left((\hat{\theta}_k - \theta)^t L (\hat{\theta}_k - \theta) \right)^p = \ell(p) \quad \text{p.s.} \quad (4.25)$$

Démonstration La preuve est donnée dans la Section 4.4 ■

Ainsi, puisque L est strictement définie positive, on déduit de (4.25) que

$$G_n(p) = \mathcal{O}(\log n) \quad \text{p.s.}$$

4.3.2 Modèles autorégressifs linéaires

On peut appliquer les résultats de la Section 4.3.1 aux modèles autorégressifs linéaires ainsi qu'aux processus de branchement avec immigration.

Le modèle autorégressif linéaire est un cas particulier du modèle de régression (4.5). Il est défini pour tout $n \geq 1$, par

$$X_{n+1} = \sum_{k=1}^d \theta_k X_{n-k+1} + \varepsilon_{n+1}. \quad (4.26)$$

La matrice compagne C associée à ce modèle est donnée par

$$C = \begin{pmatrix} \theta_1 & \theta_2 & \dots & \theta_{d-1} & \theta_d \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix},$$

Cas stable

On s'intéresse ici seulement au *cas stable*, c'est-à-dire quand $\rho(C) < 1$, où $\rho(C)$ désigne le rayon spectral de la matrice compagne C . Sous l'hypothèse C_2 , on note

$$\Gamma = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

On suppose que (ε_n) est soit une suite de variables aléatoires indépendantes et identiquement distribuées, soit une différence de martingale possédant un moment conditionnel d'ordre > 2 fini. Alors le résultat de convergence (4.24) est vérifié (voir par exemple Lai et Wei [59], Duflo [28]) avec L définie par

$$L = \sum_{k=0}^{\infty} C^k \Gamma (C^t)^k. \quad (4.27)$$

De plus, on peut facilement vérifier que L est inversible. Par conséquent, $\log d_n$ est p.s. équivalent à $d \log n$. Sous les hypothèses du corollaire 4.3.1, $\Gamma_n(2q)$ est un estimateur consistant de $\sigma(2q)$,

$$\left(\Gamma_n(2q) - \frac{1}{n} \sum_{k=1}^n \varepsilon_k^{2q} \right)^2 = \mathcal{O}\left(\frac{\log n}{n}\right) \quad \text{p.s.} \quad (4.28)$$

et le corollaire 4.3.2 s'applique en remplaçant $\log d_n$ par $d \log n$. Sous les hypothèses du corollaire 4.3.2, on a aussi

$$G_n(p) = \mathcal{O}(\log n) \quad \text{p.s.}$$

et (4.25) est également vérifiée pour L donnée en (4.27).

4.3.3 Processus de branchement avec immigration

Estimation de la moyenne

On considère le processus de branchement (X_n) sujet à une composante d'immigration indépendante à chaque génération. La notion d'immigration correspond au fait que la population de référence peut s'enrichir d'apports extérieurs : par exemple, on peut modéliser l'évolution d'un patrimoine génétique, de phénomènes en écologie, en physique des particules ou en épidémiologie. Le processus de branchement (X_n) est donné par la relation de récurrence

$$X_{n+1} = \sum_{k=1}^{X_n} Y_{n+1,k} + I_{n+1}, \quad (4.29)$$

avec $X_0 = 1$. La variable aléatoire (I_n) correspond à l'effectif de l'immigration à la génération n . Pour chaque individu k de la génération n , $Y_{n+1,k}$ désigne son nombre de descendants. On suppose que les familles de variables aléatoires (I_n) et ($Y_{n,k}$), indépendantes et identiquement distribuées, sont indépendantes entre elles. On pose alors

$$\begin{aligned}\mathbb{E}[Y_{n,k}] &\stackrel{\text{def}}{=} m, & \mathbb{E}[I_n] &\stackrel{\text{def}}{=} \lambda, \\ \text{var}[Y_{n,k}] &\stackrel{\text{def}}{=} \sigma^2, & \text{var}[I_n] &\stackrel{\text{def}}{=} b^2.\end{aligned}$$

La relation de récurrence (4.29) peut s'écrire sous la forme

$$X_{n+1} \stackrel{\text{def}}{=} mX_n + \lambda + \varepsilon_{n+1}, \quad (4.30)$$

où (ε_n) est une suite de différences de martingale. Le processus (4.29) est donc un cas particulier du modèle de régression (4.5) avec $\Phi_n^t \stackrel{\text{def}}{=} (X_n, 1)$ et $\theta^t \stackrel{\text{def}}{=} (m, \lambda)$. Cependant, dans ce modèle, le moment conditionnel d'ordre 2 du bruit n'est pas presque sûrement borné car

$$\mathbb{E}[\varepsilon_{n+1}^2 \mid \mathcal{F}_n] = \sigma^2 X_n + b^2.$$

Pour cette raison, on considère le modèle de régression suivant

$$\tilde{X}_{n+1} = \theta^t \tilde{\Phi}_n + \tilde{\varepsilon}_{n+1},$$

où les variables \tilde{X}_{n+1} , $\tilde{\Phi}_n$ et $\tilde{\varepsilon}_n$ sont définies par

$$\begin{aligned}\tilde{X}_{n+1} &\stackrel{\text{def}}{=} c_n^{-1/2} X_{n+1}, & \tilde{\Phi}_n &\stackrel{\text{def}}{=} c_n^{-1/2} \Phi_n, \\ \tilde{\varepsilon}_{n+1} &\stackrel{\text{def}}{=} c_{n+1}^{-1/2} \varepsilon_{n+1}, & c_n &\stackrel{\text{def}}{=} X_n + 1.\end{aligned}$$

Avec ce modèle, on a clairement la majoration

$$\mathbb{E}[\tilde{\varepsilon}_{n+1}^2 \mid \mathcal{F}_n] \leq \sigma^2 + b^2.$$

De plus, dans le cas stable $m < 1$, la convergence suivante a été établie par Wei et Winnicki [89]

$$\lim_{n \rightarrow \infty} n^{-1} \tilde{S}_n \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} n^{-1} \left(S + \sum_{k=0}^n c_k^{-1} \Phi_k \Phi_k^t \right) = L,$$

où L est la matrice définie positive donnée par

$$L \stackrel{\text{def}}{=} \begin{pmatrix} \mathbb{E} \left[\frac{X^2}{X+1} \right] & \mathbb{E} \left[\frac{X}{X+1} \right] \\ \mathbb{E} \left[\frac{X}{X+1} \right] & \mathbb{E} \left[\frac{1}{X+1} \right] \end{pmatrix}.$$

La notation X désigne une variable aléatoire suivant la loi stationnaire associée à (X_n) . Par conséquent, les hypothèses des corollaires 4.3.1 et 4.3.2 sont vérifiées. On obtient donc

$$\begin{aligned} \left(\Gamma_n(2q) - \frac{1}{n} \sum_{k=1}^n \varepsilon_k^{2q} \right)^2 &= \mathcal{O}\left(\frac{\log n}{n}\right) \quad \text{p.s.} \\ \left| \frac{1}{n} C_n(q) - \sigma(2q) \right|^2 &= o(n^{c-1}) \quad \text{p.s.} \end{aligned}$$

Le comportement asymptotique de l'erreur d'estimation est donné par

$$G_n(p) = \mathcal{O}(\log n) \quad \text{p.s.}$$

Estimation de la variance

La définition de (ε_n) dans la relation (4.30) permet d'écrire la décomposition

$$\varepsilon_{n+1}^2 = \sigma^2 X_n + b^2 + v_{n+1},$$

où (v_n) est une suite de différences de martingale vérifiant

$$\mathbb{E}[v_{n+1}^2 | \mathcal{F}_n] = 2\sigma^4 X_n^2 + X_n(\tau^4 - 3\sigma^4 + 4b^2\sigma^2) + \nu^4 - b^4,$$

en notant τ^4 et ν^4 les moments centrés d'ordre 4 de $(Y_{n,k})$ et (I_n) . Par conséquent, on a la majoration

$$\mathbb{E}[c_{n+1}^{-2} v_{n+1}^2 | \mathcal{F}_n] \leq \tau^4 + 4b^2\sigma^2 + \nu^4.$$

Comme la moyenne θ est inconnue, on utilise l'estimateur des variances $\eta^t \stackrel{\text{def}}{=} (\sigma^2, b^2)$ défini par

$$\widehat{\eta}_n \stackrel{\text{def}}{=} Q_n^{-1} \sum_{k=1}^n c_k^{-2} \Phi_k \widehat{\varepsilon}_{k+1}^2, \quad \widehat{\varepsilon}_{k+1} = X_{k+1} - \widehat{\theta}_k \Phi_k,$$

où

$$Q_n \stackrel{\text{def}}{=} S + \sum_{k=1}^n c_k^{-2} \Phi_k \Phi_k^t.$$

D'après Wei et Winnicki [89], dans le cas stable $m < 1$, l'hypothèse de convergence (4.24) est vérifiée puisque

$$\lim_{n \rightarrow \infty} n^{-1} Q_n = L,$$

où L est la matrice définie positive donnée par

$$L \stackrel{\text{def}}{=} \begin{pmatrix} \mathbb{E}\left[\frac{X^2}{(X+1)^2}\right] & \mathbb{E}\left[\frac{X}{(X+1)^2}\right] \\ \mathbb{E}\left[\frac{X}{(X+1)^2}\right] & \mathbb{E}\left[\frac{1}{(X+1)^2}\right] \end{pmatrix}.$$

Ainsi les hypothèses des corollaires 4.3.1 et 4.3.2 sont également vérifiées. Les comportements asymptotiques des erreurs d'estimation et de prédiction cumulées sont donc identiques au cas de l'estimation de la moyenne.

4.3.4 Un lien entre la CGR et les processus RCA

Soit $U = U_1 \dots U_n$ une suite de réalisations de variables aléatoires i.i.d. à valeurs dans l'alphabet des quatre nucléotides $\mathcal{A} = \{A, C, G, T\}$. On construit la *Chaos Game Representation* à coefficients de contraction non constants (voir Chapitre 1) de la séquence U sur le segment $[0, 1[$ et on note $(X_k)_{0 \leq k \leq n}$ les coordonnées des points obtenus. On considère ensuite que l'on n'observe qu'une version *perturbée* de la CGR

$$X_{n+1} = \sum_{u \in \mathcal{A}} \mathbb{1}_{\{U_{n+1}=u\}} (p_u X_n + \ell_u) + \varepsilon_{n+1} \quad (4.31)$$

et (ε_n) est une suite de variables aléatoires i.i.d. centrées et de variance $\sigma^2 > 0$. Ce processus correspond aux erreurs d'observation sur la CGR. Les suites (u_n) et (ε_n) sont supposées indépendantes. La relation (4.31) peut s'écrire sous la forme

$$X_{n+1} = p^t \theta_{n+1} X_n + \ell^t \theta_{n+1} + \varepsilon_{n+1}, \quad (4.32)$$

où les trois vecteurs colonnes ℓ , p et θ_{n+1} sont définis par

$$p^t \stackrel{\text{def}}{=} (p_u)_{u \in \mathcal{A}}^t, \quad \ell^t \stackrel{\text{def}}{=} (\ell_u)_{u \in \mathcal{A}}^t, \quad \text{et} \quad \theta_{n+1}^t \stackrel{\text{def}}{=} (\mathbb{1}_{\{U_{n+1}=u\}})_{u \in \mathcal{A}}^t.$$

On est donc en présence d'un processus autorégressif à coefficient aléatoire (RCA de l'anglais *Random Coefficient Autoregressive*). Supposons maintenant que la séquence U ne soit plus observable et que sa loi de génération soit inconnue. L'objectif est de retrouver cette loi, ce qui revient à estimer $\mathbb{E}[\theta_{n+1}] \stackrel{\text{def}}{=} p$. Soit $\mathbb{F} = (\mathcal{F}_n)$ la filtration naturelle du modèle avec $\mathcal{F}_n = \sigma(X_0, X_1, \dots, X_n)$. Un petit calcul permet de déterminer

$$\mathbb{E}[X_{n+1} | \mathcal{F}_n] = a X_n + \alpha, \quad (4.33)$$

$$\mathbb{E}[X_{n+1}^2 | \mathcal{F}_n] = b X_n^2 + \beta + 2\gamma X_n + \sigma^2, \quad (4.34)$$

où les paramètres a , b , c , α , β et γ sont définis par

$$\begin{aligned} a &\stackrel{\text{def}}{=} \sum_{u \in \mathcal{A}} p_u^2, & b &\stackrel{\text{def}}{=} \sum_{u \in \mathcal{A}} p_u^3, & c &\stackrel{\text{def}}{=} \sum_{u \in \mathcal{A}} p_u^4, \\ \alpha &\stackrel{\text{def}}{=} \sum_{u \in \mathcal{A}} p_u \ell_u, & \beta &\stackrel{\text{def}}{=} \sum_{u \in \mathcal{A}} p_u^2 \ell_u, & \gamma &\stackrel{\text{def}}{=} \sum_{u \in \mathcal{A}} p_u \ell_u^2. \end{aligned}$$

Notons qu'il est facile de voir que

$$\ell^t p = \frac{1 - \|p\|^2}{2}. \quad (4.35)$$

Si l'on sait estimer les trois paramètres a , b et c , qui sont en fait les sommes de Newton du polynôme de degré 4 dont les racines sont les probabilités p_u , on pourra, via les formules de Newton et les fonctions symétriques, estimer le vecteur de probabilité p . Pour tout

entier $n \geq 0$, en notant $Y_n \stackrel{\text{def}}{=} X_n - 1/2$, il est clair d'après (4.32), (4.33), (4.34) et (4.35) que

$$Y_{n+1} = a Y_n + \xi_{n+1},$$

(ξ_n) étant une suite de différences de martingale, de variance conditionnelle

$$\mathbb{E}[\xi_{n+1}^2 | \mathcal{F}_n] \stackrel{\text{def}}{=} \tau_n^2 = (b - a^2)Y_n^2 + dY_n + \delta,$$

avec $d \stackrel{\text{def}}{=} b - a^2 + 2(\gamma - a\alpha)$ et $\delta = (b - a^2)/4 + (\gamma - a\alpha) + \sigma^2 - \alpha^2$. La deuxième somme de Newton a peut s'estimer par l'estimateur des moindres carrés donné par

$$\hat{a}_n = \frac{\sum_{k=1}^n Y_{k-1} Y_k}{\sum_{k=1}^n Y_{k-1}^2}.$$

On peut montrer que \hat{a}_n converge presque sûrement vers a et établir un théorème de la limite centrale pour \hat{a}_n , ce qui fera l'objet d'un futur travail de recherche.

4.3.5 Sur l'estimateur des moindres carrés pondérés

Pour contourner certaines difficultés inhérentes au cas vectoriel, j'ai choisi dans un premier temps d'utiliser l'estimateur des moindres carrés pondérés. Ce qui va suivre résume le travail présenté dans Cénac [14]. Dans cette sous-section, et seulement ici, (M_n) désigne la transformée de martingale pondérée

$$M_n = M_0 + \sum_{k=1}^n a_{k-1} \Phi_{k-1} \varepsilon_k,$$

où a_n est une suite décroissante adaptée à la filtration \mathbb{F} , avec $0 \leq a_n \leq 1$. De plus, elle vérifie la convergence

$$\sum_{k=0}^{\infty} a_k f_k(a) < \infty \quad \text{p.s.}, \quad (4.36)$$

où le coefficient d'explosion $f_n(a)$ est défini par

$$f_n(a) \stackrel{\text{def}}{=} a_n \Phi_n^t S_n^{-1}(a) \Phi_n, \quad \text{avec} \quad S_n(a) = \sum_{k=0}^n a_k \Phi_k \Phi_k^t + S.$$

Les propriétés asymptotiques pour martingales pondérées vectorielles énoncées dans le théorème 4.3.3 permettent d'appréhender les comportements asymptotiques des erreurs cumulées de prédiction et d'estimation. Pour l'estimation de θ dans les modèles de régression linéaire présentés dans la Section 4.1.3, on choisit ici l'estimateur des moindres carrés pondérés.

Théorème 4.3.3. *On suppose que*

$$\sup_{n \geq 0} \mathbb{E}[\varepsilon_{n+1}^2 \mid \mathcal{F}_n] < \infty \quad p.s.$$

Alors, pour tout entier $p \geq 1$, on a

$$\sum_{n=1}^{\infty} \left(M_n^t S_{n-1}^{-1}(a) M_n - M_n^t S_n^{-1}(a) M_n \right)^p < \infty \quad p.s.$$

En comparaison avec le théorème 4.1.2 dans le cas scalaire ou le théorème 4.2.5, on peut noter que l'hypothèse de moment n'est vraiment pas restrictive. Grâce à la pondération, un moment conditionnel d'ordre 2 suffit. Avant d'énoncer les applications statistiques de ce théorème, on peut rappeler la définition de l'estimateur des moindres carrés pondérés

$$\hat{\theta}_n = S_{n-1}^{-1}(a) \sum_{k=1}^n a_{k-1} \Phi_{k-1} X_k. \quad (4.37)$$

Les propriétés asymptotiques de cet estimateur sont étudiées dans Bercu et Duflo [9], Bercu [7] et Guo [44]. Le choix de la pondération est crucial. Si $s_n \stackrel{\text{def}}{=} \text{tr}(S_n)$, on peut vérifier aisément que la suite définie par

$$a_n = \frac{1}{(\log s_n)^{1+\gamma}}, \quad \text{avec } \gamma > 0,$$

vérifie la convergence (4.36). Elle est un choix possible de pondération, tout comme la suite $a_n = s_n^{-\gamma}$. Les applications du théorème 4.3.3 aux erreurs cumulées sont énoncées dans les corollaires ci-dessous.

Corollaire 4.3.4. *On suppose que (s_n) est croissante vers l'infini et que*

$$\limsup_{n \rightarrow +\infty} f_n(a) < 1 \quad p.s.$$

Sous l'hypothèse (CH_{2p}) pour $p \geq 1$, avec $\mathbb{E}[\varepsilon_{n+1}^{2p} \mid \mathcal{F}_n] = \sigma(2p)$, dès que

$$a_n^{-1} = \mathcal{O}(n) \quad p.s.,$$

l'erreur de prédiction cumulée vérifie

$$\lim_{n \rightarrow \infty} \frac{C_n(p)}{n} = \sigma(2p) \quad p.s.$$

De plus, si (H_{2p+}) est vérifiée, si pour tout $n \geq 1$, $\mathbb{E}[\varepsilon_{n+1}^{2p-1} \mid \mathcal{F}_n] = 0$, et si

$$a_n^{-p} = \mathcal{O}(n^{2c}) \quad p.s.,$$

alors il existe c tel que $2pa^{-1} < c < 1$ et

$$\left| \frac{1}{n} C_n(p) - \sigma(2p) \right| = o(n^{c-1}) \quad p.s.$$

Corollaire 4.3.5. *Sous l'hypothèse (CH_{2p}) pour $p \geq 2$, on a*

$$\sum_{k=1}^{\infty} a_k f_k(a) \left((\hat{\theta}_k - \theta)^t S_k(a) (\hat{\theta}_k - \theta) \right)^p < \infty \quad p.s.$$

De plus, en supposant qu'il existe une matrice inversible L telle que

$$\lim_{n \rightarrow +\infty} \frac{1}{n} S_n = L \quad p.s.$$

il vient aussi

$$\sum_{k=1}^{\infty} k^{p-1} a_k^{p+1} \left((\hat{\theta}_k - \theta)^t L (\hat{\theta}_k - \theta) \right)^p < \infty \quad p.s.$$

et finalement

$$G_n(p) = o\left((\log s_n)^{(p+1)(1+\gamma)}\right) \quad p.s.$$

On peut appliquer ces corollaires aux modèles autorégressifs linéaires et fonctionnels.

Modèles autorégressifs linéaires

On a déjà vu dans la Section 4.3.2 que ce modèle est un cas particulier de (4.5) avec $\theta^t = (\theta_1, \dots, \theta_d)$. On s'intéresse ici au cas stable uniquement. On suppose que le bruit (ε_n) est soit une suite de variables indépendantes et identiquement distribuées, soit une suite de différences de martingale de moment conditionnel fini d'ordre supérieur à 2. De plus, on suppose qu'il existe $p \geq 2$ tel que (CH_p) soit vérifiée, alors

$$\left| \frac{1}{n} \Gamma_n(p) - \frac{1}{n} \sum_{k=1}^n \varepsilon_k^p \right|^2 = o\left(\frac{(\log n)^{1+\gamma}}{n}\right) \quad p.s.$$

Le comportement asymptotique de l'erreur d'estimation est donné par

$$G_n(p) = o\left((\log n)^{(1+\gamma)(p+1)}\right) \quad p.s.$$

Remarque 4.3.6. Si le modèle est instable, c'est-à-dire si $\rho(C) = 1$, sous l'hypothèse de moment (CH_p) avec $p > 2$, on déduit de la Proposition 4.4.24 de Duflo [28] que $f_n(a)$ converge presque sûrement vers zéro et $\log s_n = \mathcal{O}(\log n)$ p.s. Par conséquent, le corollaire 4.3.4 s'applique.

Modèles autorégressifs fonctionnels

Ces modèles sont également un cas particulier de (4.5) avec

$$\Phi_n^t = (f_1(X_n), \dots, f_d(X_{n-d+1})).$$

Pour se placer dans le cas stable, il est nécessaire d'ajouter certaines conditions restrictives. Plus précisément, on suppose que pour tout entier k tel que $1 \leq k \leq d$ et pour tout réel x ,

$$\alpha_k |x| + \beta_k \leq |f_k(x)| \leq \gamma_k |x| + \delta_k,$$

où les constantes $\alpha_k, \beta_k, \gamma_k$ et δ_k sont positives avec

$$0 < \sum_{k=1}^d \gamma_k \theta_k < 1.$$

Enfin, on suppose qu'il existe un entier k tel que $\beta_k > 0$ et un entier j tel que $\alpha_j > 0$. Dans ce contexte, avec des hypothèses de moment standards décrites ci-dessous, on peut montrer que $n = \mathcal{O}(s_n)$ p.s. et $s_n = \mathcal{O}(n)$ p.s.

Si (ε_n) est soit un bruit blanc gaussien soit une suite vérifiant, pour tout entier $n \geq 0$ et pour tout $t \geq 0$,

$$\mathbb{E}[\exp(t\varepsilon_n) | \mathcal{F}_{n-1}] \leq \exp[(\sigma^2 t^2)/2] \quad \text{p.s.},$$

alors $\Gamma_n(p)$ est un estimateur consistant de $\sigma(p)$ avec

$$\left| \frac{1}{n} \Gamma_n(p) - \frac{1}{n} \sum_{k=1}^n \varepsilon_k^p \right|^2 = o\left(\frac{(\log n)^{1+\gamma}}{n}\right) \quad \text{p.s.}$$

De plus si p est pair et que le moment conditionnel d'ordre $p-1$ est nul, alors

$$G_n(p) = o((\log n)^{(1+\gamma)(p+1)}) \quad \text{p.s.}$$

4.4 Preuves

4.4.1 Preuve du théorème 4.2.1

Démonstration Pour alléger les notations, on commence par définir les variables

$$\begin{aligned} V_n &\stackrel{\text{def}}{=} M_n^t S_{n-1}^{-1} M_n, \\ \varphi_n &\stackrel{\text{def}}{=} \alpha_n^{-1} \Phi_n^t L^{-1} \Phi_n, \\ v_n &\stackrel{\text{def}}{=} \alpha_{n-1}^{-1} M_n^t L^{-1} M_n. \end{aligned} \tag{4.38}$$

Tout d'abord, on peut remarquer que la convergence (4.12) entraîne que presque sûrement

$$f_n = \varphi_n + o(\varphi_n), \tag{4.39}$$

$$V_n = v_n + o(v_n). \tag{4.40}$$

En effet, la matrice limite L est symétrique, définie positive, elle admet donc une racine carrée inversible. Il vient alors

$$\begin{aligned} f_n &= \varphi_n + \Phi_n^t (S_n^{-1} - \alpha_n^{-1} L^{-1}) \Phi_n \\ &= \varphi_n + \alpha_n^{-1} \Phi_n^t L^{-1/2} (\alpha_n L^{1/2} S_n^{-1} L^{1/2} - I) L^{-1/2} \Phi_n. \end{aligned}$$

La matrice $R_n \stackrel{\text{def}}{=} \alpha_n L^{1/2} S_n^{-1} L^{1/2} - I$ est symétrique, donc diagonalisable dans une base orthonormale. De ce fait, si on note ρ_n son rayon spectral, on a la majoration

$$\left| \alpha_n^{-1} \Phi_n^t L^{-1/2} R_n L^{-1/2} \Phi_n \right| \leq \rho_n \varphi_n,$$

et la suite ρ_n tend vers 0 presque sûrement. Ainsi (4.39) est démontrée. On prouve (4.40) exactement de la même manière avec la décomposition

$$V_n = v_n + M_n^t (S_{n-1}^{-1} - \alpha_{n-1}^{-1} L^{-1}) M_n.$$

On déduit directement de (4.40) que $V_n^p = v_n^p + o(v_n^p)$. Pour déterminer la limite (4.13), il suffit donc, d'après le lemme de Toeplitz, d'étudier la convergence

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n f_k V_k^p = \lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n \varphi_k v_k^p. \quad (4.41)$$

On prouve le théorème 4.2.1 par récurrence sur $p \geq 1$.

Cas $p = 1$

L'idée de la preuve est inspirée de celle de Bercu [8] dans le cas scalaire. Il s'agit dans un premier temps d'écrire une relation de récurrence sur $M_n^t L^{-1} M_n$. Par définition de M_n , on peut décomposer

$$M_{n+1}^t L^{-1} M_{n+1} = M_n^t L^{-1} M_n + 2\varepsilon_{n+1} \Phi_n^t L^{-1} M_n + \varepsilon_{n+1}^2 \Phi_n^t L^{-1} \Phi_n. \quad (4.42)$$

Comme dans le cas scalaire, le terme qui donne la vitesse de convergence de $M_n^t L^{-1} M_n$ est celui qui est porté par la puissance d'ordre 2 du bruit, à savoir ici $\varepsilon_{n+1}^2 \Phi_n^t L^{-1} \Phi_n$. On définit donc

$$\beta_n \stackrel{\text{def}}{=} \sum_{k=0}^n \Phi_k^t L^{-1} \Phi_k + \text{tr}(L^{-1/2} S L^{-1/2}). \quad (4.43)$$

On pose alors $m_{n+1} \stackrel{\text{def}}{=} \beta_n^{-1} M_{n+1}^t L^{-1} M_{n+1}$. La décomposition (4.42) devient alors

$$m_{n+1} = m_n - m_n \frac{\beta_n - \beta_{n-1}}{\beta_n} + 2\varepsilon_{n+1} \beta_n^{-1} M_n^t L^{-1} \Phi_n + \varepsilon_{n+1}^2 \frac{\beta_n - \beta_{n-1}}{\beta_n}. \quad (4.44)$$

On introduit les notations supplémentaires

$$\gamma_n \stackrel{\text{def}}{=} \frac{\beta_n - \beta_{n-1}}{\beta_n}, \quad \delta_n \stackrel{\text{def}}{=} \beta_n^{-1} M_n^t L^{-1} \Phi_n.$$

En sommant l'égalité (4.44), on peut écrire que, pour tout entier $n \geq 1$,

$$m_{n+1} + \mathcal{A}_n = m_1 + \mathcal{B}_{n+1} + \mathcal{W}_{n+1}, \quad (4.45)$$

avec

$$\mathcal{A}_n \stackrel{\text{def}}{=} \sum_{k=1}^n \gamma_k m_k, \quad \mathcal{B}_{n+1} \stackrel{\text{def}}{=} 2 \sum_{k=1}^n \varepsilon_{k+1} \delta_k, \quad \mathcal{W}_{n+1} \stackrel{\text{def}}{=} \sum_{k=1}^n \varepsilon_{k+1}^2 \gamma_k.$$

En étudiant le comportement asymptotique de \mathcal{W}_{n+1} , \mathcal{B}_{n+1} et m_n , on va pouvoir établir la vitesse de convergence de \mathcal{A}_n . Il restera à comparer β_n et α_n pour en déduire la limite donnée en (4.41). Comme $\lim \alpha_n^{-1} S_n = L$, on a $\lim \alpha_n^{-1} \text{tr}(L^{-1/2} S_n L^{-1/2}) = \text{tr}(I)$, c'est-à-dire

$$d = \lim_{n \rightarrow \infty} \alpha_n^{-1} \sum_{k=0}^n \text{tr}(L^{-1/2} \Phi_k \Phi_k^t L^{-1/2}) = \lim_{n \rightarrow \infty} \alpha_n^{-1} \sum_{k=0}^n \Phi_k^t L^{-1} \Phi_k = \lim_{n \rightarrow \infty} \alpha_n^{-1} \beta_n. \quad (4.46)$$

Dans un premier temps, on peut montrer que $\mathcal{B}_{n+1} = o(\mathcal{A}_n)$. Pour ce faire, on applique la loi des grands nombres pour les séries régressives donnée par Duflo [28]. Cette loi forte entraîne que presque sûrement

$$|\mathcal{B}_{n+1}|^2 = o\left(\tau_n \log \tau_n\right) \quad \text{avec} \quad \tau_n \stackrel{\text{def}}{=} \sum_{k=1}^n \delta_k^2. \quad (4.47)$$

On peut aisément comparer le terme général de cette série avec le terme général de \mathcal{A}_n . En effet, on introduit à nouveau la racine carrée de L , et ainsi

$$\begin{aligned} \delta_n^2 &= \beta_n^{-2} M_n^t L^{-1} \Phi_n \Phi_n^t L^{-1} M_n \\ &\leq \beta_n^{-2} \lambda_{\max}(L^{-1/2} \Phi_n \Phi_n^t L^{-1/2}) M_n^t L^{-1} M_n, \end{aligned}$$

où, pour une matrice carrée A , $\lambda_{\max}(A)$ désigne la plus grande valeur propre de A . La matrice $L^{-1/2} \Phi_n \Phi_n^t L^{-1/2}$ est symétrique semi-définie positive. Sa plus grande valeur propre est donc majorée par sa trace

$$\text{tr}(L^{-1/2} \Phi_n \Phi_n^t L^{-1/2}) = \Phi_n^t L^{-1} \Phi_n = \beta_n - \beta_{n-1}.$$

Finalement, on obtient la majoration capitale pour la suite de la démonstration

$$\delta_n^2 \leq \gamma_n m_n. \quad (4.48)$$

La relation (4.47) entraîne alors

$$\mathcal{B}_{n+1} = o(\mathcal{A}_n) \quad \text{p.s.} \quad (4.49)$$

Par ailleurs, on peut montrer que

$$\lim_{n \rightarrow \infty} (\log \beta_n)^{-1} \mathcal{W}_{n+1} = \sigma^2 \quad \text{p.s.} \quad (4.50)$$

En effet, puisque le bruit (ε_n) vérifie les conditions de moment (C_2) et (H_{2+}) , d'après le lemme de Chow (voir par exemple Dufflo [28, p. 22]) on a la convergence

$$\lim_{n \rightarrow \infty} \left(\sum_{k=1}^n \gamma_k \right)^{-1} \mathcal{W}_{n+1} = \sigma^2 \quad \text{p.s.}$$

dès que la série $\sum_{k=1}^n \gamma_k$ est divergente. La suite $(\beta_n)_n$ est croissante et tend vers l'infini avec $\beta_n \sim \beta_{n-1}$. On déduit alors d'une comparaison entre série et intégrale que

$$\sum_{k=1}^n \gamma_k \sim \log \beta_n \quad \text{p.s.}$$

Donc la convergence (4.50) est prouvée. Le résultat (2.30) de Wei [91] nous permet de voir que sous l'hypothèse de moment (H_{2+}) , on a $m_n = o(\log \beta_n)$ p.s. Finalement, on a donc démontré que presque sûrement

$$\lim_{n \rightarrow \infty} \frac{1}{\log \beta_n} \sum_{k=1}^n \gamma_k m_k = \sigma^2.$$

On déduit des équivalences $\beta_n \sim d \alpha_n$, prouvée en (4.46), et $\log d_n \sim d \log \beta_n$, conséquence directe de la convergence (4.12), que

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n \varphi_k v_k = \lim_{n \rightarrow \infty} \frac{d^2}{\log d_n} \sum_{k=1}^n \gamma_k m_k = d\sigma^2 \quad \text{p.s.}$$

Grâce à l'égalité (4.41), la première partie du théorème 4.2.1 est ainsi démontré pour $p = 1$.

Pour la deuxième partie, la preuve est analogue. On écrit une relation de récurrence sur V_n . Par définition de M_n , on peut décomposer

$$\begin{aligned} V_{n+1} &= M_n^t S_n^{-1} M_n + 2\varepsilon_{n+1} \Phi_n^t S_n^{-1} M_n + \varepsilon_{n+1}^2 f_n \\ &\stackrel{\text{def}}{=} h_n + 2\varepsilon_{n+1} g_n + \varepsilon_{n+1}^2 f_n. \end{aligned}$$

En sommant cette égalité, on obtient pour tout $n \geq 1$,

$$V_{n+1} + A_n = V_1 + B_{n+1} + W_{n+1},$$

avec

$$A_n \stackrel{\text{def}}{=} \sum_{k=1}^n a_k(1), \quad B_{n+1} \stackrel{\text{def}}{=} 2 \sum_{k=1}^n \varepsilon_{k+1} g_k, \quad W_{n+1} \stackrel{\text{def}}{=} \sum_{k=1}^n \varepsilon_{k+1}^2 f_k.$$

En étudiant le comportement asymptotique de W_{n+1} , B_{n+1} et V_n , on va pouvoir établir la vitesse de convergence de A_n . Tout d'abord, on applique la loi des grands nombres pour les séries régressives, et ainsi

$$|B_{n+1}|^2 = o\left(\tilde{\tau}_n \log \tilde{\tau}_n\right) \quad \text{p.s.} \quad \text{où} \quad \tilde{\tau}_n \stackrel{\text{def}}{=} \sum_{k=1}^n g_k^2. \quad (4.51)$$

On peut comparer le terme général de la série $\tilde{\tau}_n$ avec $a_n(1)$. En effet, d'après l'égalité $S_n^{-1}\Phi_n = (1 - f_n)S_{n-1}^{-1}\Phi_n$ (Duflo [28, p. 101]) et d'après la formule de Riccati, on a

$$g_n^2 = (1 - f_n)a_n(1). \quad (4.52)$$

Finalement, on obtient

$$B_{n+1} = o(A_n) \quad \text{p.s.}$$

Le bruit (ε_n) vérifie les conditions de moment (CH_2) et (H_{2+}) , donc d'après le lemme de Chow et grâce à une comparaison entre série en intégrale qui conduit à l'équivalence $\sum_{k=1}^n f_k \sim \log d_n$, on obtient

$$\lim_{n \rightarrow \infty} (\log d_n)^{-1} W_{n+1} = \sigma^2 \quad \text{p.s.}$$

Puisque presque sûrement $m_n = o(\log \beta_n)$, l'équivalence (4.40) entraîne $V_n = o(\log d_n)$ p.s. La deuxième partie du théorème 4.2.1 est ainsi démontrée.

Remarque 4.4.1. Pour prouver le théorème 4.2.1 dans le cas $p = 1$, on se ramène finalement à l'étude de la limite

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n \gamma_k m_k^p = \frac{\ell(1)}{d^2}.$$

Dans le cas général où $p \geq 2$, on va prouver de façon analogue le lemme suivant.

Lemme 4.4.2. *Sous les hypothèses du théorème 4.2.1, on a*

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n \gamma_k m_k^p = \frac{\ell(p)}{d^{p+1}} \quad \text{p.s.} \quad (4.53)$$

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n a_k(1) m_k^{p-1} = \frac{\lambda(p)}{pd^{p-1}} \quad \text{p.s.} \quad (4.54)$$

Avant de donner la preuve du lemme 4.4.2, on peut juste préciser pourquoi le théorème 4.2.1 en est un corollaire. D'après la convergence (4.46), $\beta_n \sim d \alpha_n$, donc on a presque sûrement

$$V_n \sim d m_n, \quad \text{et} \quad f_n \sim d \gamma_n.$$

La convergence (4.53) implique alors immédiatement (4.13).

Dans le cas $p = 1$, la deuxième convergence (4.54) est exactement (4.14). Dans le cas où $p \geq 2$, il reste à comparer $a_n(1)m_n^{p-1}$ avec

$$a_n(p) \stackrel{\text{def}}{=} (M_n^t S_{n-1}^{-1} M_n)^p - (M_n^t S_n^{-1} M_n)^p.$$

On déduit de l'égalité élémentaire

$$x^p - y^p = (x - y)x^{p-1} \sum_{q=0}^{p-1} \left(\frac{y}{x}\right)^{p-1-q}, \quad (4.55)$$

appliquée à $x = V_n$ et $y = V_n - a_n(1)$, que

$$a_n(p) = a_n(1)V_n^{p-1} \sum_{q=0}^{p-1} \left(\frac{V_n - a_n(1)}{V_n}\right)^{p-1-q}.$$

En appliquant la formule de Riccati dans l'expression $a_n(1)$, on obtient

$$\begin{aligned} a_n(1) &= (1 - f_n)M_n^t S_{n-1}^{-1} \Phi_n \Phi_n^t S_{n-1}^{-1} M_n \\ &\leq (1 - f_n) \text{tr}(S_{n-1}^{-1/2} \Phi_n \Phi_n^t S_{n-1}^{-1/2}) V_n \\ &\leq f_n V_n. \end{aligned} \quad (4.56)$$

Ainsi, on a $a_n(1) = o(V_n)$ p.s. On en déduit que $a_n(p) \sim p a_n(1) V_n^{p-1}$ p.s. et finalement comme $V_n \sim d m_n$,

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n a_k(p) = p d^{p-1} \lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n a_k(1) m_k^{p-1} = \lambda(p) \quad \text{p.s.} \quad (4.57)$$

Preuve du lemme 4.4.2 par récurrence sur p

Le cas $p = 1$ est traité dans la sous-section précédente. On considère alors le cas $p \geq 2$. On suppose que le lemme 4.4.2 est vrai pour tout entier q tel que $1 \leq q \leq p - 1$.

On raisonne de la même manière que pour $p = 1$ en cherchant une relation de récurrence sur m_{n+1}^p . On déduit de (4.42) que l'on élève à la puissance p , la décomposition

$$m_{n+1}^p = \sum_{k=0}^p \sum_{\ell=0}^k 2^{k-\ell} C_p^k C_k^\ell \gamma_n^\ell \delta_n^{k-\ell} ((1 - \gamma_n) m_n)^{p-k} \varepsilon_{n+1}^{k+\ell}. \quad (4.58)$$

On fait un changement de variable dans cette double somme de façon à ordonner les termes selon les puissances du bruit (ε_{n+1}). Après quelques simplifications élémentaires on obtient une relation de récurrence de la forme

$$m_{n+1}^p + \mathcal{A}_n(p) = m_1^p + \mathcal{B}_{n+1} + \mathcal{W}_{n+1}(p), \quad (4.59)$$

avec

$$\mathcal{A}_n(p) \stackrel{\text{def}}{=} \sum_{k=1}^n \beta_k^{-p} (\beta_k^p - \beta_{k-1}^p) m_k^p, \quad \mathcal{W}_{n+1}(p) \stackrel{\text{def}}{=} \sum_{k=1}^n \gamma_k^p \varepsilon_{k+1}^{2p}, \quad \mathcal{B}_{n+1} \stackrel{\text{def}}{=} \sum_{\ell=1}^{2p-1} \sum_{k=1}^n b_k(\ell) \varepsilon_{k+1}^\ell.$$

Si $1 \leq \ell \leq p-1$, on a

$$b_k(\ell) \stackrel{\text{def}}{=} \sum_{j=0}^{\lfloor \ell/2 \rfloor} 2^{\ell-2j} C_p^{\ell-j} C_{\ell-j}^j \gamma_k^j \delta_k^{\ell-2j} ((1-\gamma_k)m_k)^{p-\ell+j},$$

tandis que si $p \leq \ell \leq 2p-1$, alors

$$b_k(\ell) \stackrel{\text{def}}{=} \sum_{j=\ell-(p-1)}^{\lfloor \ell/2 \rfloor} 2^{\ell-2j} C_p^{\ell-j} C_{\ell-j}^j \gamma_k^j \delta_k^{\ell-2j} ((1-\gamma_k)m_k)^{p-\ell+j} + C_p^{\ell-p} 2^{2p-\ell} \delta_k^{2p-\ell} \gamma_k^{\ell-p}.$$

Comme dans la preuve du cas $p=1$, on va étudier le comportement asymptotique de $\mathcal{W}_{n+1}(p)$, \mathcal{B}_{n+1} et m_n^p afin d'en déduire des informations sur $\mathcal{A}_n(p)$. Dans un premier temps, on peut montrer que $\mathcal{W}_{n+1} = o(\log d_n)$ p.s. En effet, (ε_n) vérifie la condition de moment (H_{2p+}) , donc le lemme de Chow s'applique, et

$$\mathcal{W}_{n+1} = \mathcal{O}\left(\sum_{k=1}^n \gamma_k^p\right) \text{ p.s.}$$

On déduit de l'inégalité $x \leq -\log(1-x)$, pour $0 < x < 1$, que

$$\sum_{k=1}^n \gamma_k^p \leq \sum_{k=1}^n \gamma_k \leq \sum_{k=1}^n (\log \beta_k - \log \beta_{k-1}) \leq \log \beta_n.$$

Par suite, puisque $\gamma_n \rightarrow 0$, on obtient bien $\mathcal{W}_{n+1} = o(\log d_n)$. Pour étudier \mathcal{B}_{n+1} , on centre la suite de différences de martingale (ε_n) , en posant, pour tout entier ℓ tel que $1 \leq \ell \leq 2p-1$,

$$\varepsilon_{n+1}^\ell \stackrel{\text{def}}{=} e_{n+1}(\ell) + \mathbb{E}[\varepsilon_{n+1}^\ell | \mathcal{F}_n] = e_{n+1}(\ell) + \sigma_n(\ell). \quad (4.60)$$

On décompose alors $\sum_{k=1}^n b_k(\ell) \varepsilon_{k+1}^\ell \stackrel{\text{def}}{=} \mathcal{C}_{n+1}(\ell) + \mathcal{D}_n(\ell)$, avec

$$\mathcal{C}_{n+1}(\ell) \stackrel{\text{def}}{=} \sum_{k=1}^n b_k(\ell) e_{k+1}(\ell), \quad \text{et} \quad \mathcal{D}_n(\ell) \stackrel{\text{def}}{=} \sum_{k=1}^n b_k(\ell) \sigma_k(\ell).$$

Étude de $\mathcal{C}_{n+1}(\ell)$ et $\mathcal{D}_n(\ell)$ dans le cas $1 \leq \ell \leq p-1$.

On applique à nouveau la loi des grands nombres pour les séries régressives. Pour tout entier ℓ tel que $1 \leq \ell \leq p-1$, on a presque sûrement

$$\mathcal{C}_{n+1}(\ell)^2 = \mathcal{O}(\tau_n(\ell) \log \tau_n(\ell)), \quad \text{avec} \quad \tau_n(\ell) \stackrel{\text{def}}{=} \sum_{k=1}^n b_k^2(\ell).$$

De plus, en appliquant l'inégalité (4.48), on peut écrire

$$\tau_n(\ell) = \mathcal{O}\left(\sum_{k=1}^n \gamma_k^\ell m_k^{2p-\ell}\right) = \mathcal{O}\left(\sup_{1 \leq k \leq n} m_k^p \sum_{k=1}^n \gamma_k^\ell m_k^{p-\ell}\right).$$

On déduit de la convergence (2.30) de Wei [91] que sous l'hypothèse de moment (H_{2p+}), on a $m_n^p = o((\log \beta_n)^\delta)$ p.s. avec $0 < \delta < 1$. Donc, l'hypothèse de récurrence implique alors $\tau_n(\ell) = o((\log d_n)^{1+\delta})$, et finalement

$$\mathcal{C}_{n+1}(\ell) = o(\log d_n) \quad \text{p.s.}$$

Pour l'étude de $\mathcal{D}_n(\ell)$, le cas $\ell = 2$ est un peu particulier. On le laisse de côté provisoirement. On suppose que $3 \leq \ell \leq p-1$ (ce cas ne nécessite d'être examiné que si $p \geq 4$). D'après l'inégalité de Hölder, et d'après l'hypothèse de moment, pour tout entier $1 \leq j \leq 2p-1$, les moments $|\sigma_n(j)|$ sont bornés presque sûrement, et donc

$$|\mathcal{D}_n(\ell)| = \mathcal{O}\left(\sum_{j=0}^{\lfloor \ell/2 \rfloor} \sum_{k=1}^n \gamma_k(j) |\delta_k|^{\ell-2j} m_k^{p-\ell+j}\right) = \mathcal{O}\left(\sum_{k=1}^n \gamma_k^{\ell/2} m_k^{p-\ell/2}\right),$$

la deuxième égalité étant une conséquence de (4.48). Si ℓ est pair, l'hypothèse de récurrence et la convergence presque sûre de γ_n vers 0 impliquent $\mathcal{D}_n(\ell) = o(\log d_n)$ p.s. Dans le cas où ℓ est impair, en appliquant l'inégalité de Cauchy Schwarz et l'hypothèse de récurrence on écrit

$$|\mathcal{D}_n(\ell)| = \mathcal{O}\left(\left(\sum_{k=1}^n \gamma_k m_k^{p-1}\right)^{1/2} \left(\sum_{k=1}^n \gamma_k^\ell m_k^{p-\ell}\right)^{1/2}\right) = o(\log d_n) \quad \text{p.s.}$$

On suppose maintenant que $p \leq \ell \leq 2p-1$.

Rappelons que grâce à l'inégalité de Hölder, les moments $|\sigma_n(\ell)|$ sont bornés p.s. et donc l'inégalité (4.48) implique

$$|\mathcal{D}_n(\ell)| = \mathcal{O}\left(\sum_{k=1}^n \gamma_k^{\ell/2} m_k^{p-\ell/2}\right) \quad \text{p.s.}$$

De même que dans le cas $1 \leq \ell \leq p-1$, on conclut selon la parité de ℓ , directement à partir de l'hypothèse de récurrence (si ℓ est pair et $\ell \geq 4$), ou en appliquant avant l'inégalité de Cauchy Schwarz (si ℓ est impair). Finalement, si $\ell \neq 2$, on obtient

$$\mathcal{D}_n(\ell) = o(\log d_n) \quad \text{p.s.}$$

On laisse une nouvelle fois de côté, provisoirement, le cas $\ell = 2$. Il reste à étudier la convergence des termes $\mathcal{C}_{n+1}(\ell)$. D'après le lemme de Chow, on a presque sûrement

$$\mathcal{C}_{n+1}(\ell) = o(\nu_n(\ell)) \quad \text{avec} \quad \nu_n(\ell) = \sum_{k=1}^n |b_k(\ell)|^{2p/\ell}.$$

En utilisant l'inégalité (4.48), on vérifie aisément que

$$|\nu_n(\ell)| = \mathcal{O}\left(\sum_{k=1}^n \gamma_k^p m_k^{(2p-\ell)p/\ell}\right).$$

Si $\ell > p$, on applique l'inégalité de Hölder en choisissant les exposants ℓ/p et $\ell/(\ell-p)$. On obtient alors par hypothèse de récurrence $\nu_n(\ell) = o(\log d_n)$ p.s. Pour le cas particulier $p = \ell$, on applique la loi des grands nombres pour les séries régressives. Il en découle que presque sûrement

$$|\mathcal{C}_{n+1}(\ell)|^2 = \mathcal{O}(\tau_n(p) \log \tau_n(p)) \quad \text{avec} \quad \tau_n(p) \stackrel{\text{def}}{=} \sum_{k=1}^n b_k(p)^2.$$

On déduit alors de l'inégalité (4.48) que

$$\tau_n(p) = \mathcal{O}\left(\sum_{k=1}^n \gamma_k^p m_k^p\right) = \mathcal{O}\left(\sup_{1 \leq k \leq n} m_k^p \sum_{k=1}^n \gamma_k^p\right) \quad \text{p.s.}$$

Cependant, d'après le résultat (2.30) de Wei [91], on a $m_k^p = o((\log d_n)^\delta)$, avec $0 < \delta < 1$. Donc finalement on obtient $|\mathcal{C}_{n+1}(\ell)| = o(\log d_n)$ p.s.

Il reste à traiter le cas particulier $\ell = 2$

Si $p \geq 3$, il s'agit d'étudier le comportement asymptotique de

$$\begin{aligned} \frac{\mathcal{D}_n(2)}{\log d_n} &= \frac{\sigma^2}{\log d_n} \sum_{k=1}^n \sum_{j=0}^1 2^{2-2j} C_p^{2-j} C_{2-j}^j \gamma_k^j \delta_k^{2-2j} m_k^{p-2+j}, \\ &= \frac{2p(p-1)\sigma^2}{\log d_n} \sum_{k=1}^n \delta_k^2 m_k^{p-2} + \frac{p\sigma^2}{\log d_n} \sum_{k=1}^n \gamma_k m_k^{p-1}. \end{aligned}$$

Par hypothèse de récurrence, on sait calculer la limite du deuxième terme. Pour la première somme, on va comparer δ_n^2 et g_n . En effet, on pourra alors calculer cette limite à partir de (4.52) et de l'hypothèse de récurrence pour $p-1$. Pour ce faire, on décompose $\Phi_n^t S_{n-1}^{-1} M_n$ en utilisant la racine carrée de L , comme dans la preuve de l'équivalence (4.39). On obtient

$$\begin{aligned} \Phi_n^t S_{n-1}^{-1} M_n &= \alpha_n^{-1} \Phi_n^t L^{-1} M_n + \alpha_n^{-1} \Phi_n^t (\alpha_n S_{n-1}^{-1} - L^{-1}) M_n \\ &= \alpha_n^{-1} \Phi_n^t L^{-1} M_n + \alpha_n^{-1} \Phi_n^t L^{-1/2} (\alpha_n L^{1/2} S_{n-1}^{-1} L^{1/2} - I) L^{-1/2} M_n. \end{aligned}$$

En diagonalisant la matrice $R_n = \alpha_n L^{1/2} S_{n-1}^{-1} L^{1/2} - I$ dans une base orthogonale, grâce à l'inégalité de Cauchy-Schwarz on peut écrire

$$\left| \alpha_n^{-1} \Phi_n^t L^{-1/2} R_n L^{-1/2} M_n \right| \leq \rho_n \alpha_n^{-1} (\Phi_n^t L^{-1} \Phi_n M_n^t L^{-1} M_n)^{1/2},$$

où ρ_n est le rayon spectral de R_n . Ainsi, on obtient que presque sûrement

$$g_n^2 = (\Phi_n^t S_{n-1}^{-1} M_n)^2 = \alpha_n^{-2} (\Phi_n^t L^{-1} M_n)^2 + o(\gamma_n m_n) = d^2 \delta_n^2 + o(\gamma_n m_n). \quad (4.61)$$

Finalement, on déduit du lemme de Toeplitz et de l'hypothèse de récurrence que

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n \delta_k^2 m_k^{p-2} = \lim_{n \rightarrow \infty} \frac{1}{d^2 \log d_n} \sum_{k=1}^n a_k(1) m_k^{p-2} = \frac{\lambda(p-1)}{(p-1)d^p} \quad \text{p.s.}$$

On trouve ainsi

$$\lim_{n \rightarrow \infty} \frac{\mathcal{D}_n(2)}{\log d_n} = \ell(p-1) \left(\frac{2p(p-1)\sigma^2}{d^{p+1}} + \frac{p\sigma^2}{d^p} \right) = \frac{p}{d^{p+1}} \ell(p) \quad \text{p.s.}$$

Dans le cas particulier où $p = 2$, il vient

$$\lim_{n \rightarrow \infty} \frac{\mathcal{D}_n(2)}{\log d_n} = \lim_{n \rightarrow \infty} \frac{\sigma^2}{\log d_n} \sum_{k=1}^n (4\delta_k^2 + 2\gamma_k m_k) = \ell(1)\sigma^2 \left(\frac{4}{d^3} + \frac{2}{d^2} \right) = \frac{2}{d^3} \ell(2) \quad \text{p.s.}$$

On a donc démontré que pour tout entier $\ell \in \{1, \dots, 2p-1\}$, on a presque sûrement $|\mathcal{C}_{n+1}(\ell)| = o(\log d_n)$, $|\mathcal{D}_n(\ell)| = o(\log d_n)$ sauf si $\ell = 2$, $\mathcal{W}_{n+1}(p) = o(\log d_n)$, et aussi $m_n^p = o(\log d_n)$. En reprenant la décomposition (4.59), on obtient ainsi

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \mathcal{A}_n(p) = \lim_{n \rightarrow \infty} \frac{1}{\log d_n} \mathcal{D}_n(2) = \frac{p}{d^{p+1}} \ell(p) \quad \text{p.s.}$$

La convergence (4.53) est donc démontrée car en appliquant l'égalité (4.55) à $x = \beta_{n-1}$ et $y = \beta_n$, on obtient

$$\frac{\beta_n^p - \beta_{n-1}^p}{\beta_n^p} = \gamma_n \sum_{q=0}^{p-1} \left(\frac{\beta_{n-1}}{\beta_n} \right)^{p-1-q} \sim p\gamma_n \quad \text{p.s.} \quad (4.62)$$

et donc presque sûrement

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n \gamma_k m_k^p = \lim_{n \rightarrow \infty} \frac{1}{p \log d_n} \mathcal{A}_n(p) = \frac{\ell(p)}{d^{p+1}},$$

ce qui achève la preuve de (4.53).

Pour la deuxième partie du lemme 4.4.2, on raisonne de la même façon à partir de la relation de récurrence

$$V_{n+1}^p = \sum_{k=0}^p \sum_{\ell=0}^k 2^{k-\ell} C_p^k C_k^\ell f_n^\ell g_n^{k-\ell} h_n^{p-k} \varepsilon_{n+1}^{k+\ell}, \quad (4.63)$$

où $h_n \stackrel{\text{def}}{=} M_n^t S_n^{-1} M_n$. Pour alléger les notations on définit également

$$a_n(p) \stackrel{\text{def}}{=} (M_n^t S_{n-1}^{-1} M_n)^p - (M_n^t S_n^{-1} M_n)^p.$$

On ordonne les termes selon les puissances du bruit (ε_{n+1}) et on obtient

$$V_{n+1}^p + A_n(p) = V_1^p + B_{n+1} + W_{n+1}(p), \quad (4.64)$$

avec

$$A_n(p) \stackrel{\text{def}}{=} \sum_{k=1}^n a_k(p), \quad W_{n+1}(p) \stackrel{\text{def}}{=} \sum_{k=1}^n f_k^p \varepsilon_{k+1}^{2p}, \quad B_{n+1} \stackrel{\text{def}}{=} \sum_{\ell=1}^{2p-1} \sum_{k=1}^n \tilde{b}_k(\ell) \varepsilon_{k+1}^\ell.$$

Si $1 \leq \ell \leq p-1$, on a

$$\tilde{b}_k(\ell) \stackrel{\text{def}}{=} \sum_{j=0}^{\lfloor \ell/2 \rfloor} 2^{\ell-2j} C_p^{\ell-j} C_{\ell-j}^j f_k^j g_k^{\ell-2j} h_k^{p-\ell+j},$$

tandis que si $p \leq \ell \leq 2p-1$, alors

$$\tilde{b}_k(\ell) \stackrel{\text{def}}{=} \sum_{j=\ell-(p-1)}^{\lfloor \ell/2 \rfloor} 2^{\ell-2j} C_p^{\ell-j} C_{\ell-j}^j f_k^j g_k^{\ell-2j} h_k^{p-\ell+j} + C_p^{\ell-p} 2^{2p-\ell} g_k^{2p-\ell} f_k^{\ell-p}.$$

Tout d'abord, on peut montrer que $W_{n+1} = o(\log d_n)$ p.s. En effet, le bruit (ε_n) satisfait la condition de moment (H_{2p+}) , donc le lemme de Chow s'applique, et

$$W_{n+1} = \mathcal{O}\left(\sum_{k=1}^n f_k^p\right) \quad \text{p.s.}$$

On déduit de l'inégalité $x \leq -\log(1-x)$, pour $0 < x < 1$, que

$$\sum_{k=1}^n f_k^p \leq \sum_{k=1}^n f_k \leq \sum_{k=1}^n (\log d_k - \log d_{k-1}) \leq \log d_n.$$

Par suite, puisque $f_n \rightarrow 0$, on obtient bien $W_{n+1} = o(\log d_n)$. Pour étudier B_{n+1} , on centre la suite de différences de martingale (ε_n) comme en (4.60). On décompose alors $\sum_{k=1}^n \tilde{b}_k(\ell) \varepsilon_{k+1}^\ell \stackrel{\text{def}}{=} C_{n+1}(\ell) + D_n(\ell)$, avec

$$C_{n+1}(\ell) \stackrel{\text{def}}{=} \sum_{k=1}^n \tilde{b}_k(\ell) e_{k+1}(\ell), \quad \text{et} \quad D_n(\ell) \stackrel{\text{def}}{=} \sum_{k=1}^n \tilde{b}_k(\ell) \sigma_k(\ell).$$

Étude de $C_{n+1}(\ell)$ et $D_n(\ell)$ dans le cas $1 \leq \ell \leq p - 1$.

On applique à nouveau la loi des grands nombres pour les séries régressives. Pour tout entier ℓ tel que $1 \leq \ell \leq p - 1$, on a presque sûrement

$$C_{n+1}(\ell)^2 = \mathcal{O}(\tilde{\tau}_n(\ell) \log \tilde{\tau}_n(\ell)), \quad \text{avec} \quad \tilde{\tau}_n(\ell) \stackrel{\text{def}}{=} \sum_{k=1}^n \tilde{b}_k^2(\ell).$$

On déduit aisément de (4.52) et de l'inégalité (4.56) que

$$g_n^2 \leq f_n V_n \tag{4.65}$$

De plus, puisque $h_n \leq V_n$, on a

$$\tilde{\tau}_n(\ell) = \mathcal{O}\left(\sum_{k=1}^n f_k^\ell V_k^{2p-\ell}\right) = \mathcal{O}\left(\sup_{1 \leq k \leq n} V_k^p \sum_{k=1}^n f_k^\ell V_k^{p-\ell}\right).$$

La convergence (2.30) de Wei [91] implique que sous l'hypothèse de moment (H_{2p+}), on a $V_n^p = o((\log d_n)^\delta)$ p.s. avec $0 < \delta < 1$. Donc, l'hypothèse de récurrence implique alors $\tilde{\tau}_n(\ell) = o((\log d_n)^{1+\delta})$, et finalement

$$C_{n+1}(\ell) = o(\log d_n) \quad \text{p.s.}$$

Pour l'étude de $D_n(\ell)$, on suppose dans un premier temps que $3 \leq \ell \leq p - 1$. Pour tout entier $1 \leq j \leq 2p - 1$, les moments $|\sigma_n(j)|$ sont bornés presque sûrement, et donc

$$|D_n(\ell)| = \mathcal{O}\left(\sum_{j=0}^{\lfloor \ell/2 \rfloor} \sum_{k=1}^n f_k(j) |g_k|^{\ell-2j} V_k^{p-\ell+j}\right) = \mathcal{O}\left(\sum_{k=1}^n f_k^{\ell/2} V_k^{p-\ell/2}\right),$$

la deuxième égalité étant une conséquence de (4.65). Si ℓ est pair, l'hypothèse de récurrence et la convergence presque sûre de f_n vers 0 impliquent $D_n(\ell) = o(\log d_n)$ p.s. Dans le cas où ℓ est impair, en appliquant l'inégalité de Cauchy Schwarz et l'hypothèse de récurrence on écrit

$$|D_n(\ell)| = \mathcal{O}\left(\left(\sum_{k=1}^n f_k V_k^{p-1}\right)^{1/2} \left(\sum_{k=1}^n f_k^\ell V_k^{p-\ell}\right)^{1/2}\right) = o(\log d_n) \quad \text{p.s.}$$

On suppose maintenant que $p \leq \ell \leq 2p - 1$.

Les moments $|\sigma_n(\ell)|$ sont bornés p.s. et donc on déduit de (4.65) que

$$|D_n(\ell)| = \mathcal{O}\left(\sum_{k=1}^n f_k^{\ell/2} V_k^{p-\ell/2}\right) \quad \text{p.s.}$$

On conclut selon la parité de ℓ directement à partir de l'hypothèse de récurrence (si ℓ est pair et $\ell \geq 4$), ou en appliquant avant l'inégalité de Cauchy Schwarz (si ℓ est impair). Finalement, si $\ell \neq 2$, on obtient

$$D_n(\ell) = o(\log d_n) \quad \text{p.s.}$$

On laisse une nouvelle fois de côté, provisoirement, le cas $\ell = 2$. Il reste à étudier la convergence des termes $C_{n+1}(\ell)$. D'après le lemme de Chow, on a presque sûrement

$$C_{n+1}(\ell) = o(\tilde{\nu}_n(\ell)) \quad \text{avec} \quad \tilde{\nu}_n(\ell) = \sum_{k=1}^n |\tilde{b}_k(\ell)|^{2p/\ell}.$$

En utilisant l'inégalité (4.65), on vérifie aisément que

$$|\tilde{\nu}_n(\ell)| = \mathcal{O}\left(\sum_{k=1}^n f_k^p V_k^{(2p-\ell)p/\ell}\right).$$

Si $\ell > p$, on applique l'inégalité de Hölder en choisissant les exposants ℓ/p et $\ell/(\ell-p)$. On obtient alors par hypothèse de récurrence $\tilde{\nu}_n(\ell) = o(\log d_n)$ p.s. Pour le cas particulier $p = \ell$, on applique la loi des grands nombres pour les séries régressives. Il en découle que presque sûrement

$$|C_{n+1}(\ell)|^2 = \mathcal{O}(\tilde{\tau}_n(p) \log \tilde{\tau}_n(p)) \quad \text{avec} \quad \tilde{\tau}_n(p) \stackrel{\text{def}}{=} \sum_{k=1}^n \tilde{b}_k(p)^2.$$

On déduit alors de l'inégalité (4.65) que

$$\tilde{\tau}_n(p) = \mathcal{O}\left(\sum_{k=1}^n f_k^p V_k^p\right) = \mathcal{O}\left(\sup_{1 \leq k \leq n} V_k^p \sum_{k=1}^n f_k^p\right) \quad \text{p.s.}$$

Rappelons que d'après le résultat (2.30) de Wei [91], on a $V_k^p = o((\log d_n)^\delta)$, avec $0 < \delta < 1$. Donc finalement on obtient $|C_{n+1}(\ell)| = o(\log d_n)$ p.s.

Il reste à traiter le cas particulier $\ell = 2$

Si $p \geq 3$, il s'agit d'étudier

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{D_n(2)}{\log d_n} &= \frac{\sigma^2}{\log d_n} \sum_{k=1}^n \sum_{j=0}^1 2^{2-2j} C_p^{2-j} C_{2-j}^j f_k^j g_k^{2-2j} V_k^{p-2+j}, \\ &= \lim_{n \rightarrow \infty} \frac{2p(p-1)\sigma^2}{\log d_n} \sum_{k=1}^n g_k^2 V_k^{p-2} + \lim_{n \rightarrow \infty} \frac{p\sigma^2}{\log d_n} \sum_{k=1}^n f_k V_k^{p-1}. \end{aligned}$$

Par hypothèse de récurrence, on sait calculer la limite du deuxième terme. Pour la première somme, on déduit de (4.52), de l'équivalence presque sûre $V_n \sim d m_n$ et de l'hypothèse de récurrence pour $p - 1$ que

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n g_k^2 V_k^{p-2} = \lim_{n \rightarrow \infty} \frac{d^{p-2}}{\log d_n} \sum_{k=1}^n a_k(1) m_k^{p-2} = \frac{\lambda(p-1)}{p-1} \quad \text{p.s.}$$

On trouve ainsi

$$\lim_{n \rightarrow \infty} \frac{D_n(2)}{\log d_n} = p\sigma^2 \left(2\lambda(p-1) + \ell(p-1) \right) = \lambda(p) \quad \text{p.s.}$$

Dans le cas particulier où $p = 2$, il vient

$$\lim_{n \rightarrow \infty} \frac{D_n(2)}{\log d_n} = \lim_{n \rightarrow \infty} \frac{\sigma^2}{\log d_n} \sum_{k=1}^n (4g_k^2 + 2f_k V_k) = \sigma^2 \left(4\lambda(1) + 2\ell(1) \right) = \lambda(2) \quad \text{p.s.}$$

On a donc démontré que pour tout entier $\ell \in \{1, \dots, 2p-1\}$, on a presque sûrement $|C_{n+1}(\ell)| = o(\log d_n)$, $|D_n(\ell)| = o(\log d_n)$ sauf si $\ell = 2$, $W_{n+1}(p) = o(\log d_n)$, et aussi $V_n^p = o(\log d_n)$. En reprenant la décomposition (4.64), on obtient ainsi

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} A_n(p) = \lim_{n \rightarrow \infty} \frac{1}{\log d_n} D_n(2) = \lambda(p) \quad \text{p.s.}$$

Il reste à appliquer l'égalité des limites dans l'équation (4.57) pour achever la démonstration du lemme 4.4.2. ■

4.4.2 Preuve du corollaire 4.2.4

Démonstration Dans le cas particulier $p = 1$, la convergence (4.18) est exactement (4.14). On suppose maintenant $p > 1$. On déduit de l'inégalité (4.56) et de la convergence presque sûre de f_n vers zéro, en appliquant le lemme de Kronecker et le lemme 4.4.2 que

$$0 \leq \lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n (a_k(1))^p \leq \lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n (f_k V_k)^p = 0 \quad \text{p.s.}$$

■

Le corollaire 4.2.4 est donc démontré.

4.4.3 Preuve du théorème 4.2.5

Comme dans la preuve du théorème 4.2.1, il suffit de démontrer le lemme suivant.

Lemme 4.4.3. *Sous les hypothèses du théorème 4.2.5, on a*

$$\sum_{k=1}^n \gamma_k m_k^p = \mathcal{O}(\log d_n) \quad \text{p.s.} \quad (4.66)$$

Démonstration On prouve le lemme 4.4.2 par récurrence sur p .

Cas $p = 1$

On utilise à nouveau la décomposition (4.45). Les arguments qui ont permis d'écrire $m_n = o(\log d_n)$, $\mathcal{B}_{n+1} = o(\mathcal{A}_n)$ sont toujours valables sous les hypothèses du théorème 4.2.5. De plus, d'après le lemme de Chow, on a $\mathcal{W}_{n+1} = \mathcal{O}(\log d_n)$ p.s. Par conséquent, on obtient bien $\mathcal{A}_{n+1} = \mathcal{O}(\log d_n)$ p.s. et (4.66) est démontrée pour $p = 1$.

Réurrence sur p .

Dans l'égalité (4.59), on a déjà prouvé dans le théorème 4.2.5 que $m_{n+1} = o(\log d_n)$, $\mathcal{W}_{n+1} = o(\log d_n)$ et $\mathcal{C}_{n+1}(\ell) = o(\log d_n)$ p.s. pour tout entier $\ell \in \{1, \dots, 2p - 1\}$ car les arguments utilisés sont encore valables sous les hypothèses du théorème 4.2.5. Avec l'inégalité de Hölder et l'hypothèse (H_{2p+}) , on peut toujours écrire $|\sigma_n(\ell)| = \mathcal{O}(1)$, pour tout entier $\ell \geq 2$, et ainsi on a encore $\mathcal{D}_n(\ell) = o(\log d_n)$ pour tout $\ell \geq 3$. Il reste à traiter le cas $\ell = 2$. Or, puisque les moments conditionnels d'ordre 2 sont bornés p.s., on a

$$\mathcal{D}_n(2) = \mathcal{O}\left(\sum_{k=1}^n \sum_{j=0}^1 4^{1-j} C_p^{2-j} C_{2-j}^j \gamma_k^j \delta_k^{2-2j} m_k^{p-2+j}\right) \quad \text{p.s.}$$

et grâce à la majoration (4.48), on peut écrire

$$\mathcal{D}_n(2) = \mathcal{O}\left(\sum_{k=1}^n \gamma_k m_k^{p-1}\right) = \mathcal{O}(\log d_n) \quad \text{p.s.},$$

la deuxième égalité étant donnée par l'hypothèse de récurrence. Finalement, on obtient $\mathcal{A}(p) = \mathcal{O}(\log d_n)$ p.s., ce qui compte tenu de l'équivalence (4.62), permet de conclure la preuve du lemme 4.4.3. ■

4.4.4 Preuve du corollaire 4.3.1

Démonstration En développant l'expression de $\Gamma_n(2q)$, (4.16) implique

$$n(\Gamma_n(2q) - \Delta_n(2q)) = \sum_{k=0}^n \pi_k^{2q} + \sum_{\ell=1}^{2q-1} C_q^\ell \sum_{k=0}^{n-1} \pi_k^{2q-\ell} \varepsilon_{k+1}^\ell.$$

D'après (4.17), $\pi_k^2 \sim a_k(1)$, on déduit du corollaire 4.2.4

$$\sum_{k=0}^n \pi_k^{2q} = o(\log d_n) \quad \text{si } q > 1, \quad \text{et} \quad \sum_{k=0}^n \pi_k^2 = \mathcal{O}(\log d_n) \quad \text{p.s.}$$

De plus, pour tout $\ell \in \{1, \dots, 2q - 1\}$, on décompose $\sum_{k=0}^{n-1} \pi_k^{2q-\ell} \varepsilon_{k+1}^\ell$ en deux termes,

$$\sum_{k=0}^{n-1} \pi_k^{2q-\ell} \varepsilon_{k+1}^\ell = \sum_{k=0}^{n-1} \pi_k^{2q-\ell} e_{k+1}(\ell) + \sum_{k=0}^{n-1} \pi_k^{2q-\ell} \sigma_k(\ell).$$

Rappelons que les moments sont presque sûrement bornés et ainsi grâce à (4.48), on a

$$\sum_{k=0}^{n-1} \pi_k^{2q-\ell} \sigma_k(\ell) = \mathcal{O}\left(\sum_{k=0}^{n-1} \gamma_k^{2q-\ell} m_k^{2q-\ell}\right) = \mathcal{O}(\log d_n) \quad \text{p.s.}$$

Finalement, d'après le lemme de Chow, on peut écrire

$$\left| \sum_{k=0}^{n-1} \pi_k^{2q-\ell} e_{k+1}(\ell) \right| = o\left(\sum_{k=1}^{n-1} \pi_k^{2q-\ell}\right) = o(\log d_n) \quad \text{p.s.}$$

Par conséquent, on obtient

$$n^2 (\Gamma_n(2q) - \Delta_n(2q))^2 = \mathcal{O}(n \log d_n) \quad \text{p.s.},$$

ce qui achève la preuve du corollaire 4.3.1. ■

4.4.5 Preuve du corollaire 4.3.2

Démonstration On montre facilement à partir de la définition de S_n et de l'estimateur des moindres carrés $\hat{\theta}_n$ que

$$(\hat{\theta}_n - \theta)^t S_n (\hat{\theta}_n - \theta) = V_n + g_n^2.$$

Ainsi, on déduit de la convergence (4.13), de la limite nulle du coefficient d'explosion f_n et du lemme de Kronecker que presque sûrement

$$\lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n f_k \left((\hat{\theta}_k - \theta)^t S_k (\hat{\theta}_k - \theta) \right)^p = \lim_{n \rightarrow \infty} \frac{1}{\log d_n} \sum_{k=1}^n f_k V_k^p = \ell(p).$$

Ainsi la convergence (4.23) est une conséquence immédiate du théorème 4.2.1. Pour la deuxième partie du corollaire, on peut déjà montrer en utilisant la racine carrée de L que

$$\left((\hat{\theta}_n - \theta)^t L (\hat{\theta}_n - \theta) \right)^p \sim (n^{-2} m_n \beta_n)^p \sim (d^2 m_n \beta_n^{-1})^p \quad \text{p.s.}$$

Ainsi, on obtient

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n k^{p-1} \left((\hat{\theta}_k - \theta)^t L (\hat{\theta}_k - \theta) \right)^p = \lim_{n \rightarrow \infty} \frac{d^{p+1}}{\log n} \sum_{k=1}^n \frac{m_k^p}{\beta_k} \quad \text{p.s.}$$

On cherche à comparer cette somme avec celle étudiée dans le lemme 4.4.2. Grâce à une transformation d'Abel, on a la décomposition

$$\sum_{k=1}^n \gamma_k m_k^p = \frac{m_n^p}{\beta_n} (\Sigma_n - d(n-1)) - \frac{m_1^p}{\beta_0} \Sigma_0 + r_n + d \sum_{k=1}^{n-1} \frac{m_k^p}{\beta_k}, \quad (4.67)$$

avec

$$\Sigma_n \stackrel{\text{def}}{=} \sum_{k=1}^n \beta_k \gamma_k = \sum_{k=1}^n \Phi_k^t L^{-1} \Phi_k \sim \beta_n$$

et

$$r_n \stackrel{\text{def}}{=} \sum_{k=1}^{n-1} \left(\frac{m_k^p}{\beta_k} - \frac{m_{k+1}^p}{\beta_{k+1}} \right) (\Sigma_k - kd).$$

On peut voir aisément que

$$\frac{m_n^p}{\beta_n} (\Sigma_n - d(n-1)) - \frac{m_1^p}{\beta_0} \Sigma_0 = o(\log d_n) \quad \text{p.s.}$$

Il reste à prouver que $r_n = o(\log d_n)$ p.s. En effet, sous cette hypothèse, le lemme 4.4.2 implique

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n \frac{m_k^p}{\beta_k} = \frac{1}{d} \lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n \gamma_k m_k^p = \frac{\ell(p)}{d^{p+1}} \quad \text{p.s.}$$

Pour prouver que r_n tend vers zéro presque sûrement, on décompose r_n en deux termes de la façon suivante,

$$r_n = \sum_{k=1}^{n-1} \frac{\Sigma_k - kd}{\beta_k} (m_k^p - m_{k+1}^p) + \sum_{k=1}^{n-1} \frac{\Sigma_k - kd}{\beta_k} \gamma_{k+1} m_{k+1}^p. \quad (4.68)$$

En utilisant la preuve du théorème 4.2.1 et la décomposition (4.58), on trouve pour le premier terme

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^{n-1} \frac{\Sigma_k - kd}{\beta_k} \left(\beta_k^{-p} (\beta_k^p - \beta_{k-1}^p) m_k^p - w_{k+1} - b_{k+1} \right) = 0 \quad \text{p.s.}$$

On déduit directement du lemme 4.4.2 et de la convergence presque sûre de $\Sigma_k - kd/\beta_k$ vers zéro que le deuxième terme est aussi un $o(\log d_n)$. Finalement, on obtient

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n k^{p-1} \left((\hat{\theta}_k - \theta)^t L (\hat{\theta}_k - \theta) \right)^p = \ell(p) \quad \text{p.s.}$$

■

Conclusions et perspectives

De nombreuses perspectives s'inscrivent dans la continuité des travaux présentés dans cette thèse. Ces dernières pages ont vocation à proposer quelques pistes de recherche.

Les nouvelles applications de la CGR proposées dans les deux premiers chapitres permettent de comparer des séquences biologiques et de tirer avantage des propriétés intrinsèques de la construction. Cependant, le problème lié à la détermination de distances pertinentes biologiquement, entre deux mesures empiriques résultant de la CGR, n'est pas complètement résolu. On peut également penser à des moyens de comparaison non asymptotiques.

D'autre part, les études sont ici présentées dans le cadre de représentation CGR avec coefficient de contraction ρ constant. Qu'en est-il des représentations introduites par Gutiérrez et al. [45] où ce paramètre bouge avec le temps ?

Dans le chapitre 2, le problème d'un choix rigide de partition est souligné. Il est contourné par la généralisation du test à une collection de partitions. Une étude asymptotique plus fine sur la puissance du test permettrait de mieux appréhender l'influence du choix. En effet, les expérimentations numériques semblent indiquer une dépendance certaine entre vitesse de convergence de la puissance et partition choisie pour le test.

Le comportement asymptotique des branches critiques dans les arbres-CGR est identique, au premier ordre, à celui des arbres digitaux de recherche construits à partir d'insertion de séquences indépendantes. Il serait intéressant d'obtenir le deuxième terme dans le développement. On peut par exemple penser aux *tries* construits à partir de suffixes : Fayolle [35] montre que ce n'est qu'à partir du deuxième ordre que le comportement de la profondeur d'insertion diffère des *tries* classiques.

La construction des arbres-CGR était motivée par la volonté de mesurer de nouvelles quantités statistiques, cachées dans la séquence mais visibles sur l'arbre, afin de dégager de nouvelles caractéristiques pour une loi de génération donnée. Un théorème de la limite centrale sur les trois statistiques considérées dans le chapitre 3, à savoir les longueurs extrêmes des branches et la profondeur d'insertion, est un prolongement naturel des convergences obtenues ici. Ce résultat est associé au comportement asymptotique des variables au deuxième ordre.

Enfin, une autre piste de recherche serait de s'intéresser au cas de sources dynamiques, au lieu de chaînes de Markov. Le comportement asymptotique au premier ordre serait-il identique ?

Dans le chapitre 4, l'hypothèse de convergence sur le processus croissant

$$\lim_{n \rightarrow \infty} \alpha_n^{-1} S_n = L \quad \text{p.s.}$$

où L est une matrice inversible, n'est certainement pas optimale pour garantir les convergences du théorème 4.2.1. Cette hypothèse permet de contourner les difficultés inhérentes au cas vectoriel, en se rapportant à une vitesse de convergence scalaire commune. On pourrait penser à des hypothèses moins restrictives, proches de celles de Wei [91].

En réduisant les hypothèses du théorème 4.2.1 et en l'étendant au cas explosif où le coefficient d'explosion f_n ne tend pas vers zéro presque sûrement mais vers une variable aléatoire f telle que $0 < f < 1$, on pourrait augmenter le spectre des applications statistiques en considérant par exemple des processus autorégressifs ou des processus de branchements multitypes explosifs.

Table des figures

1.1	Exemple de CGR : représentation de ATGCGAGTGT	3
1.2	Exemples de CGR : représentations de 70 000 nucléotides d' <i>Homo Sapiens</i> et de <i>Streptomyces Coelicolor</i>	3
1.3	CGR du mot ATGCGAGTGT sur le segment unité	5
1.4	CGR classique sur le tétraèdre d'une séquence i.i.d.	5
1.5	Nouvelle CGR sur le tétraèdre de séquence d' <i>Homo Sapiens</i>	7
1.6	Définition des carrés correspondant aux nucléotides et aux dinucléotides pour la CGR sur $[0, 1]^2$	7
1.7	Propriétés autosimilaires des nouvelles CGR définies sur le segment et dans le tétraèdre	8
1.8	Chemin d'intégration Γ	17
1.9	Chemins d'intégrations \mathcal{C} et \mathcal{L}	18
2.1	Les quatre différentes partitions du carré unité choisies pour le test	30
2.2	Probabilité d'acceptation de H_m en fonction de m pour <i>Homo Sapiens</i> et <i>Mus musculus</i>	36
2.3	Probabilité d'acceptation de H_m en fonction de m et de la longueur n pour des séquences d' <i>Homo Sapiens</i> et pour plusieurs partitions	38
2.4	Probabilité d'acceptation de H_m en fonction de m pour plusieurs espèces	39
2.5	Arbre taxonomique communément admis des espèces utilisées dans les simulations.	49
2.6	Arbres taxonomiques non enracinés construits avec la méthode <i>Neighbor-Joining</i> à partir des différences d'abondance relative	55
2.7	Arbre taxonomique construit à partir des différences d'abondance relative avec la partitions régulière de taille 10×10	56
2.8	Arbres taxonomiques construits à partir des différences d'abondance relative pour les séquences de bactéries	57
2.9	Arbre taxonomique construit à partir des différences d'abondance relative avec une partitions de 400 zones régulières regroupées aléatoirement en 16 ensembles.	58
2.10	Arbre taxonomique construit à partir des différences d'abondance relative, correspondant au comptage des trinuécléotides.	59
2.11	Capture d'écran du programme <code>mycgr.x</code>	64
3.1	Étapes successives de construction d'un arbre-CGR	68

3.2	Représentation de GAGCACAGTGGGAAGGG dans l'arbre-CGR et dans le carré unité	71
3.3	CGR et représentation de l'arbre sans étiquettes de 400 000 nucléotides d' <i>Homo Sapiens</i>	71
3.4	Importance de la structure de recouvrement	78
3.5	Convergence dynamique des longueurs de branches et de la profondeur d'insertion en fonction de n pour une séquence i.i.d. équiprobable ou non équiprobable	94
3.6	Histogramme des hauteurs, des longueurs de plus courtes branches et des profondeurs d'insertion	95
3.7	Capture d'écran du programme <code>mycgr.x</code> pour la partie arbres-CGR. . . .	97

Liste des tableaux

2.1	Taux de rejet de H_0	32
2.2	Taux de rejet de H_0 pour plusieurs collections de partitions arbitraires . .	33
2.3	Taux de rejet de H_1	35
2.4	Liste des séquences biologiques utilisées dans les tests	35
2.5	Liste des séquences biologiques utilisées pour la signature génomique . .	48
2.6	Différences d'abondance relative basée sur la CGR construites en utilisant la partition \mathcal{P}_2	51
2.7	Différences d'abondance relative de dinucléotides	52
2.8	Différences d'abondance relative basées sur la CGR construites à partir d'une grille régulière de taille 10×10 formant une partition	53
2.9	Liste des séquences supplémentaires utilisées dans la deuxième série d'ex- périences pour la signature génomique	56

Bibliographie

- [1] Milton Abramowitz et Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964. ISBN 0-486-61272-4.
- [2] D. Aldous et P. Shields. A diffusion limit for a class of randomly-growing binary search trees. *Probab. Theory Related Fields*, 79 :509–542, 1988.
- [3] J.S. Almeida, J.A. Carrigo, A. Maretzek, P.A. Noble, et Fletcher M. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics*, 17(5) :429–437, 2001.
- [4] V. Anh, K.S. Lau, et Z-G. Yu. Multifractal characterisation of complete genomes. *Physica A*, 301(1-4) :351–361, 2001.
- [5] Y. Baraud, S. Huet, et B. Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1) :225–251, 2003. ISSN 0090-5364.
- [6] M. T. Barlow, R. Pemantle, et E. A. Perkins. Diffusion-limited aggregation on a tree. *Probab. Theory Related Fields*, 107(1) :1–60, 1997. ISSN 0178-8051. URL <http://www.math.upenn.edu/pemantle/papers/perkins.pdf>.
- [7] B. Bercu. Weighted estimation and tracking for ARMAX models. *SIAM J. Control Optimization*, 33(1) :89–106, 1995.
- [8] B. Bercu. On the convergence of moments in the almost sure central limit theorem for martingales with statistical applications. *Stochastic Processes and their applications*, 111 :157–173, 2004.
- [9] B. Bercu et M. Duflo. Moindres carrés pondérés et poursuite. *Ann. Inst. Henri Poincaré*, 28(3) :403–430, 1992.
- [10] P. Billingsley. *Probability and Measure*. Wiley Series in Probability & Mathematical Statistics : Probability and Mathematical Statistics, 1995.
- [11] G. Blom et D. Thorburn. How many random digits are required until given sequences are obtained? *Journal of Applied Probabilities*, 19 :518–531, 1982.
- [12] G. A. Brosamler. An almost everywhere central limit theorem. *Math. Proc. Cambridge Philos. Soc.*, 104(3) :561–574, 1988. ISSN 0305-0041.

- [13] A.M. Campbell, J. Mrázek, et S. Karlin. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA*, 96 : 9184–9189, 1999.
- [14] P. Cénac. Almost sure properties of weighted vectorial martingales transforms with applications to prediction for linear regression models. *Probab. Math. Statist.*, 23 (1, Acta Univ. Wratislav. No. 2539) :61–76, 2003. ISSN 0208-4147.
- [15] P. Cénac. Test on the structure of biological sequences via chaos game representation. *Stat. Appl. Genet. Mol. Biol.*, 4 :Art. 27, 36 pp. (electronic), 2005. ISSN 1544-6115.
- [16] P. Cénac, B. Chauvin, N. Pouyanne, et Ginouillac S. Digital search trees and chaos game representation. Technical Report 5856, INRIA, march 2006.
- [17] P. Cénac, G. Fayolle, et J.M. Lasgouttes. Dynamical systems in the analysis of biological sequences. Technical Report 5351, INRIA, october 2004.
- [18] F. Chaâbane. Version forte du théorème de la limite centrale fonctionnel pour les martingales. *C. R. Acad. Sci. Paris Sér. I Math.*, 323(2) :195–198, 1996. ISSN 0764-4442.
- [19] F. Chaâbane. Invariance principles with logarithmic averaging for martingales. *Studia Sci. Math. Hungar.*, 37(1-2) :21–52, 2001. ISSN 0081-6906.
- [20] F. Chaâbane et F. Maâouia. Théorèmes limites avec poids pour les martingales vectorielles. *ESAIM Probab. Statist.*, 4 :137–189 (electronic), 2000. ISSN 1292-8100.
- [21] F. Chaâbane, F. Maâouia, et A. Touati. Généralisation du théorème de la limite centrale presque-sûr pour les martingales vectorielles. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(2) :229–232, 1998. ISSN 0764-4442.
- [22] G.A. Churchill. Stochastic models for heterogeneous dna sequences. *Bull. Math. Biol.*, 51(1) :79–94, 1989.
- [23] D. Dacunha-Castelle et M. Duflo. *Probabilités et statistiques*. Masson, 1982.
- [24] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, et Fertil B. Genomic signature : Characterization and classification of species assessed by Chaos Game Representation of sequences. *Mol. Bio. Evol.*, 16 :1391–1399, 1999.
- [25] L. Devroye et R. Neininger. Random suffix search trees. *Random Structures Algorithms*, 23(4) :357–396, 2003. ISSN 1042-9832.
- [26] J. Dieudonné. *Calcul Infinitésimal*. Hermann, 1997.

- [27] M. Drmota. The variance of the height of digital search trees. *Acta Informatica*, 38 :261–276, 2002.
- [28] M. Duflo. *Random Iterative Methods*. Springer-Verlag, 1997.
- [29] M. Duflo, R. Senoussi, et A. Touati. Propriétés asymptotiques presque sûres de l'estimateur des moindres carrés d'un modèle autoregressif vectoriel. *Ann. Inst. Henri Poincaré*, 27(1) :1–25, 1991.
- [30] M. El Karoui, V. BiauDET, S. Schbath, et A. Gruss. Characteristics of chi distribution on several bacterial genomes. *Research in Microbiology*, 150 :579–587, 1999.
- [31] J. Ellson, E. Gansner, Y Koren, E. Koutsofios, J. Mocenigo, S. North, et G. Woodhull. Graphviz - graph visualization software. <http://www.graphviz.org/>, 2005.
- [32] P. Erdős et P. Révész. On the length of the longest head run. In I. Csizàr et P. Elias, editors, *Topics in Information Theory*, volume 16, pages 219–228, North-Holland, Amsterdam, 1975. Colloq. Math. Soc. János Bolyai.
- [33] P. Erdős et P. Révész. On the length of the longest head-run. In *Topics in information theory (Second Colloq., Keszthely, 1975)*, pages 219–228. Colloq. Math. Soc. János Bolyai, Vol. 16. North-Holland, Amsterdam, 1977.
- [34] K. J. Falconer. *Fractal Geometry : Mathematical Foundations and Applications*. J. Wiley and sons, 1990.
- [35] J. Fayolle. *Compression de données sans perte et combinatoire analytique*. PhD thesis, Université Paris VI, 2006.
- [36] B. Fertil, M. Massin, S. Lespinats, C. Devic, P. Dumeé, et A. Giron. Genstyle : exploration and analysis of dna sequences with genomic signature. *Nucleic Acids Research*, 0(2) :512–515, 2005. ISSN 0305-1048.
- [37] J.C. Fu. Bounds for reliability of large consecutive-k-out-of-n :f system. *IEEE trans. Reliability*, (35) :316–319, 1986.
- [38] J.C. Fu et M.V. Koutras. Distribution theory of runs : a markov chain approach. *J. Amer. Statist. Soc.*, (89) :1050–1058, 1994.
- [39] H. Gerber et S. Li. The occurrence of sequence patterns in repeated experiments and hitting times in a markov chain. *Stochastic Processes and their Applications*, (11) :101–108, 1981.
- [40] N. Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res.*, 21 (10) :2487–2491, 1993.

- [41] G.C. Goodwin et K.S. Sin. *Adaptative Filtering Prediction and Control*. Prentice-Hall, Englewood Cliffs, N.J., 1984.
- [42] L. Gordon, M.F. Schilling, et M.S. Waterman. An extreme value theory for long head runs. *Probability Theory and related Fields*, (72) :279–287, 1986.
- [43] I.S. Gradshteyn et I.M Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 1980.
- [44] L. Guo. Self-convergence of weighted least-squares with applications to stochastic adaptive control. *IEEE Trans. Automatic Control*, 41(1) :79–89, 1996.
- [45] J.M. Gutiérrez, M.A. Rodríguez, et G. Abramson. Multifractal analysis of DNA sequences using a novel Chaos Game Representation. *Physica A*, 300 :271–284, 2001.
- [46] D. Hall et C. Heyde. *Martingale Limit Theory and its Applications*. Academic press, 1980.
- [47] H.J. Jeffrey. Chaos Game Representation of gene structure. *Nucleic Acid. Res*, 18 : 2163–2170, 1990.
- [48] R.W. Jernigan et RH. Baran. Pervasive properties of the genomic signature. *BMC Genomics*, 3(23), 2002.
- [49] J. Josse, A.D. Kaiser, et A. Kornberg. Enzymatic synthesis of deoxyribonucleic acid. VIII. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem*, 263 :864–875, 1961.
- [50] S. Karlin et C. Burge. Dinucleotide relative abundance extremes : a genomic signature. *Trends Genet.*, 7 :283–290, 1995.
- [51] S. Karlin, I. Landunga, et B.E. Blaisdell. Heterogeneity of genomes : measures and values. *Proc Natl Acad Sci USA*, 91 :12837–12841, 1994.
- [52] S. Karlin et J. Mrázek. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA*, 94 :10227–10232, 1997.
- [53] S. Karlin et J. Mrázek. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA*, 95 :3720–3725, 1998.
- [54] S. Karlin, J. Mrázek, et A.M. Campbell. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, 179(12) :3899–3913, 1997.
- [55] D. E. Knuth. *The art of computer programming. Vol. 2*. Addison-Wesley Publishing Co., Reading, Mass., second edition, 1981. ISBN 0-201-03822-6. Seminumerical algorithms, Addison-Wesley Series in Computer Science and Information Processing.

- [56] M. V. Koutras. Waiting times and number of appearances of events in a sequence of discrete random variables. In *Advances in combinatorial methods and applications to probability and statistics*, Stat. Ind. Technol., pages 363–384. Birkhäuser Boston, Boston, MA, 1997.
- [57] M. Lacey. Laws of the iterated logarithm for partial sum processes indexed by functions. *J. Theoret. Probab.*, 2(3) :377–398, 1989. ISSN 0894-9840.
- [58] T.L. Lai et C.Z. Wei. Least-squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1) :154–166, 1982.
- [59] T.L. Lai et C.Z. Wei. Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *Journal of Multivariate Analysis*, 13 :1–23, 1983.
- [60] P. L’Ecuyer. Uniform random number generation. *Ann. Oper. Res.*, 53 :77–120, 1994. ISSN 0254-5330. Simulation and modeling.
- [61] X. Leroy, D. Doligez, J. Garrigue, D. Rémy, et J. Vouillon. The Objective Caml system, documentation and user’s manual. <http://caml.inria.fr/>, 2005.
- [62] S.-Y. R. Li. A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Ann. Probab.*, 8(6) :1171–1176, 1980. ISSN 0091-1798.
- [63] M. A. Lifshits. Lecture notes on almost sure limit theorems. *Publications IRMA*, 54 :1–25, 2001.
- [64] M. A. Lifshits. Almost sure limit theorem for martingales. In *Limit theorems in probability and statistics, Vol. II (Balatonlelle, 1999)*, pages 367–390. János Bolyai Math. Soc., Budapest, 2002.
- [65] Michel Loève. *Probability theory. II*. Springer-Verlag, New York, fourth edition, 1978. ISBN 0-387-90262-7. Graduate Texts in Mathematics, Vol. 46.
- [66] H. Mahmoud. *Evolution of Random Search Trees*, chapter 6. John Wiley, New York, 1992.
- [67] H. M. Martinez. An efficient method for finding repeats in molecular sequences. *nucacires*, 11(13) :4629–4634, 1983.
- [68] A. J. Menezes, P. C. van Oorschot, et S. A. Vanstone. *Handbook of applied cryptography*. CRC Press Series on Discrete Mathematics and its Applications. CRC Press, Boca Raton, FL, 1997. ISBN 0-8493-8523-7. With a foreword by Ronald L. Rivest.

- [69] S. P. Meyn et R. L. Tweedie. *Markov chains and stochastic stability*. Springer, 1993.
- [70] F. Muri. Modelling bacterial genomes using hidden markov models. In R. Payne et P.J. Green, editors, *Compstat'98 Proceedings in Computational Statistics*, pages 89–100, Heildeberg, 1998. Physica-Verlag.
- [71] W Penney. Problem : Penney-ante. *J. Recreational Math.*, 2 :241, 1969.
- [72] G. Perrière et M. Gouy. www-query : An on-line retrieval system for biological sequence banks. *Biochimie.*, 78 :364–369, 1996.
- [73] V. Petrov. On the probabilities of large deviations for sums of independent random variables. *Theory Prob. Appl.*, (10) :287–298, 1965.
- [74] B. Pittel. Asymptotic growth of a class of random trees. *Annals Probab.*, 13 : 414–427, 1985.
- [75] V. Pozdnyakov, J. Glaz, M. Kulldorff, et J. M. Steele. A martingale approach to scan statistics. *Ann. Inst. Statist. Math.*, 57(1) :21–37, 2005. ISSN 0020-3157.
- [76] M. Régnier. A unified approach to word occurrence probabilities. *Discrete Applied Mathematics*, 104 :259–280, 2000.
- [77] G. Reinert, S. Schbath, et M.S. Waterman. Probabilistic and statistical properties of words : An overview. *Journal of Computational Biology*, 7(1/2) :1–46, 2000.
- [78] S. Robin et J.J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.*, 36 :179–193, 1999.
- [79] E. Rocha, A. Viari, et A. Danchin. Oligonucleotide bias in *Bacillus subtilis* : general trends and taxonomic comparisons. *Nucleic Acids Research*, 26 :2971–2980, 1998.
- [80] A. Roy, C. Raychaudhury, et A. Nandy. Novel techniques of graphical representation and analysis of DNA sequences – a review. *J. Biosci.*, 23(1) :55–71, 1998.
- [81] G.J. Russel et J.H. Subak-Sharpe. Similarity of the general designs of protochordates and invertebrates. *Nature(London)*, 266 :533–535, 1977.
- [82] N. Saitou et M. Nei. The neighbor-joining method : A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*, 4(4) :406–425, 1987.
- [83] S.S. Samarova. On the length of the longest head-run for a markov chain with two states. *Theory of probability and its applications*, 26(3) :498–509, 1981.
- [84] P. Schatte. On strong versions of the central limit theorem. *Math. Nachr.*, 137 : 249–256, 1988. ISSN 0025-584X.

- [85] P. Schatte. On the central limit theorem with almost sure convergence. *Probab. Math. Statist.*, 11(2) :237–246 (1991), 1990. ISSN 0208-4147.
- [86] D. Stark. First occurrence in pairs of long words : a Penney-ante conjecture of Pevzner. *Combin. Probab. Comput.*, 4(3) :279–285, 1995. ISSN 0963-5483.
- [87] V. Stefanov et A. G. Pakes. Explicit distributional results in pattern formation. *Annals of Applied Probabilities*, 7 :666–678, 1997.
- [88] A.W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
- [89] C. Z. Wei et J. Winnicki. Estimation of the means in the branching process with immigration. *Ann. Statist.*, 18(4) :1757–1773, 1990.
- [90] C.Z. Wei. Asymptotic properties of least-squares estimates in stochastic regression models. *The Annals of Statistics*, 13(4) :1498–1508, 1985.
- [91] C.Z. Wei. Adaptative prediction by least squares predictors in stochastic regression models with applications to time series. *The Annals of Statistics*, 15(4) :1667–1682, 1987.
- [92] P. Weiner. Linear pattern matching algorithm. In *14th Annual IEEE Symposium on Switching and Automata Theory*, pages 1–11, Washington, DC, 1973.
- [93] D. Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. ISBN 0-521-40455-X ; 0-521-40605-6.
- [94] J. Ziv et A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Information Theory*, IT-23(3) :337–343, 1977. ISSN 0018-9448.

Étude statistique de séquences biologiques et convergence de martingales

Résumé : La *Chaos Game Representation* (CGR) est un système dynamique qui, à une séquence de lettres dans un alphabet fini, fait correspondre une trajectoire dans un espace continu, voire une mesure empirique sur un ensemble. La CGR fournit-elle plus d'information que les méthodes de comptage de mots classiques ? On montre qu'il est possible, à partir d'une caractérisation basée sur la CGR, de déterminer un modèle de structure de séquence biologique. On construit alors une nouvelle famille de tests caractérisant l'ordre d'une chaîne de Markov homogène.

On propose ensuite une construction d'arbres digitaux de recherche, inspirés par la CGR, en insérant successivement tous les *préfixes retournés* d'une séquence générée par une chaîne de Markov d'ordre un. On montre que les longueurs des branches critiques dans de tels arbres se comportent, au premier ordre, comme si les séquences insérées formaient des chaînes de Markov indépendantes les unes des autres.

La dernière partie est consacrée à l'étude de nouvelles propriétés de convergence presque sûre de martingales vectorielles. On montre en particulier que, sous certaines conditions de régularité du processus croissant, il y a convergence des moments normalisés de tout ordre dans le théorème de la limite centrale presque sûr. Les résultats de convergence sont appliqués aux modèles de régression linéaire, ainsi qu'aux processus de branchement, afin d'établir de nouvelles propriétés asymptotiques sur les erreurs d'estimation et de prédiction.

Mots-clés : système de fonction itérée, test du chi-deux, signature génomique, arbre aléatoire, arbre digital de recherche, profondeur d'insertion, martingales, lois fortes.

Statistical investigation of biological sequences and convergence of martingales

Abstract : The *Chaos Game Representation* (CGR) is a dynamical system which maps a sequence of letters taken from a finite alphabet onto a continuous space, even an empirical measure on a set. We show how the CGR can be used to characterize the distribution of a sequence. This allows to define a new family of tests, giving the order of an homogeneous Markov chain.

In a second part, we propose a construction of *Digital Search Trees* (DST), inspired from the CGR, by successively inserting all the returned prefixes of a Markov chain. We give the asymptotic behavior of the critical lengths of paths, which turns out to be, at first order, the same one as in the case of DST built from independent Markov chains.

A last part deals with properties of almost sure convergence of vectorial martingales. In particular, under suitable regularity conditions on the growing process, we establish the convergence of normalized moments of all orders in the almost sure central limit theorem. The results are applied to linear regression models, as well as branching processes, in order to get asymptotic properties on the cumulated errors of estimation and prediction.

Key-words : iterated function system, chi-squared test, genomic signature, random tree, digital search tree, insertion depth, martingales, strong laws.