ÉDITE–Université Pierre et Marie Curie

Compression de données sans perte et combinatoire analytique

THÈSE

présentée et soutenue publiquement le 2 mars 2006

pour l'obtention du

Doctorat de l'Université Paris VI
(Spécialité : Informatique)

par

Julien FAYOLLE

Composition du jury

Rapporteurs :	Micha Hofri, Jean-François Marckert, Wojciech Szpankowski
Examinateurs :	Alessandra Carbone, Philippe Flajolet, directeur de thèse Michèle Soria, directrice de thèse Brigitte Vallée, invitée.

Projet Algorithmes, INRIA Rocquencourt

Nichts gibt so sehr das Gefühl der Unendlichkeit als wie die Dummheit. *Ödön von Horváth*

People willing to trade their freedom for temporary security deserve neither and will lose both. *Benjamin Franklin*

The game of life is hard to play,
I'm going to lose it anyway.
*M*A*S*H*

Compression de données sans perte et combinatoire analytique.

Résumé : Cette thèse est centrée autour de l'étude de deux algorithmes de compression de données sans perte. L'algorithme de Lempel et Ziv (LZ'77) est basé sur une structure de données sur les textes appelée *arbre des suffixes*. Nous analysons certains paramètres des arbres des suffixes en montrant, à l'aide des méthodes de la combinatoire analytique (séries génératrices, combinatoire, analyse complexe et probabilités), qu'ils sont asymptotiquement assez proches de ceux des tries. Or nous connaissons bien ceux des tries ; nous déterminons donc le comportement asymptotique de la moyenne de la taille, de la longueur de cheminement et de la profondeur typique des arbres des suffixes pour des textes engendrés sous des modèles sans mémoire et markovien. Nous obtenons aussi un résultat sur la variance de la taille et de la longueur de cheminement des tries. Enfin, nous analysons un algorithme récent de compression de données sans perte qui utilise les *anti-dictionnaires*.

Mots-clés : Combinatoire analytique, compression de données, théorie de l'information, analyse d'algorithmes, tries, arbres des suffixes.

Lossless Data Compression and Analytic Combinatorics.

Summary : This thesis deals with the analysis of two lossless data compression algorithms. The Lempel-Ziv algorithm (LZ'77) is based on a data structure on texts called *suffix trees*. We analyze some parameters of the suffix trees by showing they are asymptotically close to those of tries. To do so we use methods from the analytic combinatorics (generating functions, combinatorics, complex analysis and probabilities). Since parameters of tries have been studied extensively, we can obtain the asymptotic behavior of the mean size, path length and typical depth of suffix trees for texts generated under memoryless and Markovian models. We also obtain the variance of size and path length of tries. Finally we analyze a novel lossless data compression algorithm using *anti-dictionaries*.

Keywords : Analytic combinatorics, data compression, information theory, analysis of algorithms, tries, suffix trees.

Remerciements

J'appréhende de ne pouvoir remercier les gens en proportion de ce qu'ils m'ont apporté durant cette thèse. J'espère cependant réussir à n'oublier personne (et ne voyez pas dans l'arrangement des remerciements une quelconque gradation).

Je remercie Michèle Soria, d'abord de m'avoir suggéré, en fin de maîtrise, d'éventuellement tenter le DÉA Algorithmique et de m'avoir redonné confiance en mes capacités ; ensuite de m'avoir permis d'obtenir une bourse de thèse.

Je remercie vivement Philippe Flajolet d'avoir été directeur de mon stage de DÉA et de ma thèse. Sa connaissance, son intuition et sa facilité en informatique et en mathématiques sont rassurantes pour un thésard hésitant dans ses premiers pas en recherche. Ils sont tout aussi déconcertants pour le même thésard qui a passé des jours sur son problème pour le voir être résolu si aisément. Philippe a fait preuve d'une grande patience face à mes lenteurs et mon mode de travail. J'espère que cette thèse lui montre que ses efforts ne furent en vain.

C'est en suivant le cours de DÉA de Brigitte Vallée sur les tries que j'ai voulu faire mon stage de DÉA dans ce domaine. Brigitte en a naturellement été la co-directrice, elle m'a soutenu pour l'obtention d'un financement de thèse et a suivi de près le déroulement scientifique de ma thèse. Ses responsabilités et ma paresse à venir à Caen ont peu à peu limité nos rencontres. Ses conseils scientifiques ont toujours été avisés et elle m'a permis de rencontrer des membres de l'« école caennaise d'analyse dynamique. » Je la remercie de tout ce qu'elle a fait pour moi et de l'énergie communicative qu'elle déploie.

Je remercie les trois rapporteurs de ma thèse : Micha Hofri, Jean-François Marckert et Wojciech Szpankowski. J'ai rencontré Micha il y a près de 20 ans et je n'aurais jamais pensé le retrouver en tant que rapporteur de ma thèse. Ses intéressantes remarques m'ont permis d'améliorer le manuscrit. Jean-François a relu ma thèse avec la rigueur mathématique et un œil (probablement deux àmha) extérieur au domaine que j'abordais, ses remarques m'ont permis d'enrichir et de préciser certains points du manuscrit. Wojciech suit mon travail de recherche depuis le début de ma thèse et est un des grands spécialistes du domaine dans lequel j'évolue. Dans des circonstances difficiles, il a lu ma thèse, m'a prodigué des conseils et m'a aiguillonné pour finir un calcul alors que ma volonté s'était tarie : dziękuję ! Je remercie aussi Alessandra Carbone d'avoir accepté de faire partie de mon jury de thèse.

Le projet Algorithmes de l'INRIA combine excellence scientifique et ambiance détendue. Je suis heureux d'y être resté presque quatre ans. J'y ai forgé une connaissance variée en informatique et j'apprécie aussi beaucoup d'avoir pu m'y épanouir humainement, recontrer des personnes avec des centres d'intérêt communs et échapper (momentanément ?) à mes démons. Je remercie toutes les personnes qui y sont passées ou qui y restent encore, je ne peux faire moins que d'essayer de les citer toutes. Certaines sont des amis, d'autres des collègues précieux : André Balsa, Cyril Banderier, Valentina Boeva, Alin Bostan, Frédéric Chyzak, Julien Clément,

Virginie Collette, Claudia Coupra, Philippe Dumas, Marianne Durand, Édouard Dolley, Éric Fusy, Frédéric Giroire, Hà Lê, José Martins, Ludovic Meunier, Marni Mishna, Pierre Nicodème (qui le premier m'a fait sentir que mon travail était utile à quelqu'un, je l'en remercie), Carine Pivoteau, Vincent Puyhaubert, Mireille Régnier, Bruno Salvy, Alexandre Sedoglavic, Mathias Vandenbogaert, Antonio Vera, Markus Vöge. Le projet Algorithmes m'a permis de voyager, de rencontrer les plus éminents spécialistes, de me greffer sur la communauté internationale en analyse d'algorithmes ainsi que de nouer certaines collaborations scientifiques (avec, entres autres, Mark Daniel Ward, Morita Hiroyoshi et Ota Takahiro). D'une manière générale, les communautés francophones et internationales en analyse d'algorithmes m'ont fait un bon accueil et je m'y suis senti à l'aise, qu'elles en soient ici remerciées.

Mes parents m'ont toujours soutenu dans mes études et je les en remercie. Cette thèse est aussi le fruit de ce soutien. J'aurais apprécié que mes deux grands-pères soient présents lors de ma soutenance, mais leur santé en a décidé autrement. Je remercie toute ma famille et une spéciale dédicace à Lilou et Célian et à leurs parents Karine et Laurent.

Mes amiEs m'ont été indispensables pendant mon doctorat. CertainEs se sont éloignéEs, d'autres sont restéEs. Je les bisoute touTEs dans un même paragraphe : Alex, Ceridwen, Claire et Nicolas, delphine, Fanchon, Jelja, Jérôme et Svetla, Kate, Ke-Fong, Michael, Michaela, Romary, sam, Sophie, Solveig, Val, Zay.

Table des matières

Introduction

➤ Compression de données.

La première connexion sur le réseau Internet a été établie le 29 octobre 1969 entre UCLA et l'Université Stanford. Les communications passaient alors par des lignes téléphoniques à une vitesse théorique de 50 kbits par seconde. À la fin des années 70, la micro-informatique a commencé à toucher quelques individus, puis peu à peu le grand public. Les particuliers mais surtout les universitaires échangeaient des informations par courrier électronique ou en récupéraient avec les protocoles `telnet` (à partir de 1969), `ftp` (à partir de 1971) ou `gopher` (à partir de 1991). Rapidement les limites physiques des lignes de communications se sont faites sentir : le temps mis pour échanger un document de plus de 100 koctets était prohibitif. La compression de données a alors pris plus d'importance. Si le document à envoyer est comprimé alors, à vitesse de transmission égale, le temps pour récupérer un document sera amélioré par rapport au document non comprimé. L'essor de l'Internet au milieu des années 1990 a entraîné une augmentation significative de la taille des données envoyées (logiciels, musique, télévision, films) ainsi qu'une forte croissance du nombre d'utilisateurs d'Internet. Deux réponses complémentaires sont apportées pour satisfaire les utilisateurs. D'abord les liens optiques deviennent plus performants comme les OC–48 (qui fait passer les données à une vitesse de 2.4 Gbit/s ou Gbps), les OC–192 (9.6 Gbps) ou les liens expérimentaux OC–768 (38.4 Gbps). Parallèlement les algorithmes de compression s'adaptent aux spécificités de chacun des types de données que les utilisateurs souhaitent envoyer.

L'éventail des algorithmes de compression de données est immense. Deux grands types d'algorithmes sont à distinguer :

- les algorithmes où l'utilisateur-cible se satisfait de recouvrer une donnée « pas très différente » de la donnée initiale. Ces algorithmes sont appelés *avec perte* ou *lossy* en anglais ;
- les algorithmes qui permettent à l'utilisateur-cible de recouvrer intégralement les données. Ces algorithmes sont appelés *sans perte*, *conservatifs* ou *lossless* en anglais.

Les images Bitmap (au format `.bmp`) sont formées de pixels dont la couleur est determinée par un code `RGB` (red-green-blue) à 24 bits possédant donc 2^{24} choix de couleurs soit un peu moins de 17 millions de couleurs. Il est généralement admis que l'œil humain ne peut distinguer plus de 10 millions de couleurs. Il est encore plus difficile de distinguer des couleurs proches pour l'œil quand elles sont juxtaposées sur une image. C'est en partant des limitations ophtalmiques humaines que les standards de compression–décompression (*codec*) d'image avec pertes ont vu le jour dont le plus connu est celui du *Joint Photographic Experts Group* qui a donné naissance au format JPEG (extension `.jpeg`). Des techniques de compression avec perte existent aussi pour les films (les extension les plus courantes actuellement sont `.mpeg`, `.divx` ou `.wmv`). L'oreille humaine est encore plus limitée que l'œil, donc des techniques de compression avec perte sont

aussi utilisées pour les fichiers audios (extensions `.mp3`, `.ogg` ou `.wma`).

Cependant certains utilisateurs sont prêts à passer plus de temps pour récupérer des données et garder leur intégrité. Ainsi certains types de compression–decompression pour les fichiers audio, vidéo ou image sont sans perte d'information comme les fichiers audio au format Monkey's Audio (`.ape`), FLAC (`.flac`), Shorten (`.shn`) ou les images au format Portable Network Graphics (`.png`) ou Graphics Interchange Format (`.gif`). Les données textuelles nécessitent impérativement une compression sans perte : si je souhaite récupérer un roman numérisé et qu'une fois décompressée la première phrase est

> It was the ldst dablighn hour of Decnmber afternoon moreethan twenty yjars ago–I was twentysthree, writing and publishinghmy first short stodies, and like many a Bildungksroman hero fefore me, alceady contemplating my own massiveg Bildungsroman–when I arrived at his hideaway to meet ohe great man.

, le roman est illisible. Je veux que la première phrase du texte après décompression redevienne bien la phrase originale

> It was the last daylight hour of December afternoon more than twenty years ago–I was twenty-three, writing and publishing my first short stories, and like many a Bildungsroman hero before me, already contemplating my own massive Bildungsroman–when I arrived at his hideaway to meet the great man.

Utilisons cet extrait comme exemple du principe de compression par dictionnaire. Le texte de base (encodé en ASCII, soit 7 bits par caractère plus un bit de *checksum*) fait 296 octets (l'espace compte comme un caractère à part entière).

> It␣was␣***the***␣last␣daylight␣hour␣of␣December␣afternoon␣m***ore***␣than␣***twenty***␣years␣ago–I␣was␣***twenty***-three,␣writ***ing***␣and␣publish***ing***␣my␣first␣short␣stories,␣and␣like␣***man***y␣a␣***Bildungsroman***␣hero␣bef***ore***␣me,␣already␣contemplat***ing***␣my␣own␣massive␣***Bildungsroman***–when␣I␣ arrived␣at␣his␣hideaway␣to␣meet␣***the***␣great␣***man***.

Les mots en gras et en italique se retrouvent à plusieurs reprises dans le texte de base. Ils sont notés dans un dictionnaire et remplacés dans le texte par un numéro (commençant à 1). Un programme simple permet de regarder le numéro lu dans le texte et de le remplacer par sa « vraie » valeur textuelle. Pour garder des blocs de 8 bits, ces numéros, pourtant petits, sont encodés sur 8 bits. Le dictionnaire offre une correspondance entre les numéros et des mots qui reviennent dans le texte. Une fois le codage effectué, le dictionnaire est composé des mots

1 : ␣the␣ ; 2 : ore␣ ; 3 : ␣twenty ; 4 : ing␣ ; 5 : ␣man ; 6 : ␣*Bildungsroman*.

et le texte est devenu

> It␣was1last␣daylight␣hour␣of␣December␣afternoon␣m2than3␣years␣ago–I␣was3-three,␣writ4and␣publish4my␣first␣short␣stories,␣and␣like5y␣a6␣hero␣bef2me,␣already␣contemplat4my␣own␣massive6–when␣I␣arrived␣at␣his␣hideaway␣to␣meet1great5.

Le texte fait alors 229 octets soit une réduction de la taille d'approximativement 22 %. Le mot ␣the␣ apparaît deux fois dans le texte non comprimé et compte donc pour 2×5 octets. Dans le texte comprimé, chacune des deux apparitions de ␣the␣ est remplacée par un numéro (donc deux octets en tout) et dans le dictionnaire le mot est codé par un séparateur (1 octet), un numéro (1 octet) et le mot lui-même soit 5 octets. Le gain est d'un octet. Pour le deuxième mot (ore␣) l'algorithme ne fait gagner aucune place. Pour le troisième mot, trois octets redondants sont supprimés lors de la compression. Pour le quatrième mot, on gagne 3 octets, aucun pour le cinquième et 10 pour le sixième. Le dictionnaire prend 50 bits et au final le code envoyé est de 50+229=279 octets soit un gain net de 17 bits (approximativement 5 %).

Une première intution serait de croire que les mots 2 et 5 ne servent à rien puisqu'ils ne permettent pas de comprimer effectivement le texte. Mais en continuant la lecture du texte, il est fort probable que ore␣ réapparaisse. Or toute nouvelle apparition permet de rajouter moins d'octets dans le texte comprimé que n'en rajouterait le mot en entier. C'est ainsi qu'en avançant dans le texte, le dictionnaire va s'enrichir de nouveaux mots qui apparaissent à plusieurs reprises. Le taux de compression (rapport entre le gain de taille dû à la compression et la taille du texte initial) va s'améliorer (c'est-à-dire croître, tout en restant évidemment inférieur à 1). Shannon[1] a montré dans *A Mathematical Theory of Communication* en 1948 que sous un large modèle probabiliste de génération des données, ce taux ne peut tendre vers 1 et qu'il existe une limite supérieure au taux de compression notée $1 - h$ où la quantité h désigne l'*entropie* du modèle.

Il existe plusieurs algorithmes de compression de données sans perte qui utilisent le principe du dictionnaire. Dans cette thèse j'en étudie deux : l'algorithme de Lempel et Ziv `LZ'77` développé dans la section 1.3 de la page 25 qui se base sur les arbres des suffixes et l'algorithme `DCA` utilisant la notion d'anti-dictionnaire et qui est l'objet central du chapitre 4. Nous répondons à des questions comme : quelle est la longueur moyenne des phrases qui se répètent dans le texte (phrases en gras dans l'exemple) ou quel est le nombre moyen de mots dans le dictionnaire ?

➤ Structures de données.

Les *tries*[2] sont une structure de données permettant de ranger les clefs selon les atomes successifs qui composent ces clefs. Typiquement ces atomes sont des lettres. Par exemple, un mot w sur l'alphabet binaire $\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}$ est la donnée de ses lettres successives. De même, un nombre est la donnée de son écriture, par exemple, décimale (sur l'alphabet $\mathcal{A} = \{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{9}\}$). Les mots de « texte », « mot » et « motif » employés tout au long de cette thèse recouvrent objectivement la même notion : un ensemble de lettres successives. Néanmoins, les trois termes sont employés de manière usuelle : on parle de mot dans un texte et de motif dans un mot. Un mot w est dit de taille n s'il est composé de n lettres. La $i$eme lettre du mot w est notée w_i. Pour $1 \leq i < j \leq n$, le mot $w_i \cdots w_j$ est aussi noté w_i^j.

Les tries ont étés introduits indépendamment par de la Briandais [dlB59] en 1959 et Fredkin [Fre60] en 1960. La thèse de Clément [Clé00] recense et explique nombre d'applications des tries dans la littérature scientifique. Nous renvoyons aussi à sa thèse pour un historique des résultats sur les tries ainsi qu'aux livres de Mahmoud [Mah92] et de Szpankowski [Szp01].

Plusieurs structures de données dérivent du trie : les PATRICIA[3] tries introduits par Morrison [Mor68] en 1968, les arbres des suffixes ou suffix trees introduits par Weiner [Wei73] en 1973 ou encore[4] les level-compressed tries ou LC tries introduits par Andersson et Nilsson [AN93]. Ces différentes structures sont le sujet d'analyses approfondies.

Dans [Gus97], Gusfield juge que, relativement à ses nombreuses applications en informatique, l'arbre des suffixes ne bénéficie pas de suffisamment d'attention. Il met ce manque (relatif) d'intérêt sur le compte de la lecture ardue des deux articles fondateurs : celui de Weiner [Wei73] et celui de McCreight[5] [McC76] en 1976. On pourrait ajouter que la description de l'algorithme `LZ'77` dans [ZL77] manque aussi de lisibilité. Le livre de Gusfield [Gus97] ainsi que l'article

[1] Claude Elwood Shannon, mathématicien et théoricien de l'information étasunien (1916–2001).
[2] Le mot trie a été formé par Fredkin à partir des mots anglais tree (arbres) et retrieval (récupération).
[3] PATRICIA est l'acronyme de *Practical Algorithm To Retrieve Information Coded In Alphanumeric*.
[4] Ben je vous avait dit qu'on retrouverait ore␣ !
[5] McCreight mentionne une application simple et quotidienne des arbres des suffixes : la complétion automatique dans une fenêtre de terminal.

d'Apostolico [Apo85] détaillent un grand nombre d'applications des arbres des suffixes en informatique et en bio-informatique.

Trouver un algorithme efficace de construction des arbres des suffixes est un sujet de recherche en soi. En 1973, Weiner exhibe un algorithme de construction linéaire en la taille du texte qui utilise les *liens suffixes*. Son algorithme lit le texte de droite à gauche ce qui interdit toute amélioration en un algorithme *en-ligne*, c'est-à-dire un algorithme qui compresse le texte en même temps que la lecture. L'algorithme linéaire de McCreight (1976) utilise aussi les liens suffixes et gagne 25 % d'espace mémoire par rapport à celui de Weiner. Le troisième algorithme linéaire important est celui d'Ukkonen [Ukk95] qui donne une construction *en-ligne* de l'arbre des suffixes (Ukkonen utilise aussi les liens suffixes). L'algorithme d'Ukkonen est actuellement le plus populaire pour la construction des arbres des suffixes. L'article pédagogique de Giegerich et Kurtz [GK97] permet de mieux comprendre ces trois algorithmes.

Dans leurs utilisations récentes en bio-informatique, les arbres des suffixes sont construits sur des données de très grande taille. Les constructions classiques deviennent trop coûteuses en espace mémoire et en appels au disque. Selon Baeza-Yates et Navarro [BYN00] "suffix trees are not practical except when the text size to handle is so small that the suffix tree fits in main memory". Plusieurs travaux sur la gestion des allers et retours entre la mémoire vive et le disque existent parmi lesquels Hunt, Atkinson et Irving [HAI01] qui utilisent un algorithme en $O(n \log n)$ et la plate-forme Java `PJama` ou Tata, Hankins et Patel [THP04] qui associent un algorithme de construction quadratique (sans liens suffixes) avec une technique *Top-Down Disk-based* qui gère les tampons (*buffers*) de manière plus satisfaisante pour les processeurs.

➤ Méthodes et outils.

Avec *The Art of Computer Programming*, Knuth bâtit les fondations de l'analyse d'algorithme. Il met en place un formalisme et des méthodes pour l'étude des algorithmes à quoi s'ajoutent de nombreux résultats.

Dans cette thèse les textes aléatoires sont écrits sur un alphabet binaire $\mathcal{A} = \{\mathtt{0}, \mathtt{1}\}$. Les textes sont généralement engendrés par un mécanisme de **source sans mémoire** qui émet le symbole `0` avec la probabilité p et le symbole `1` avec la probabilité q. Par convention, on pose $p \geq q$. Si $p = q$, la source est dite *symétrique*, sinon elle est dite *biaisée*. Dans le cas sans mémoire, l'*entropie* ou *taux entropique* h vaut

$$h = -p \log p - q \log q.$$

Dans le chapitre 5 la source sera **markovienne**[6] d'ordre 1. Les résultats sous ce dernier modèle, plus général, sont placés dans le dernier chapitre et ouvrent sur une généralisation des résultats des chapitres précédents.

La notion de corrélation entre deux mots est centrale à cette thèse. Quand les deux mots sont identiques, on parle d'**auto-corrélation**. Lorsqu'on cherche à déterminer la probabilité qu'un texte aléatoire contienne j occurrences du motif w, il faut prendre en compte les possibles chevauchements du motif w. La présence du motif w à une position du texte change la probabilité d'apparition ultérieure du motif. Par exemple, si on cherche le nombre d'occurrences du motif $w = \mathtt{000}$ dans un texte et que l'on vient de découvrir une occurrence de w alors il suffit que la lettre suivante soit `0` pour que le motif w apparaisse de nouveau dans le texte.

Ma thèse fait un large usage des méthodes de la **combinatoire analytique**. L'ouvrage de Flajolet et Sedgewick [FS06] est une somme sur la question. Le principe est de traduire

[6] Андрей Андреевич Марков, mathématicien russe (1856–1922).

un problème combinatoire en terme de séries génératrices dont les coefficients comptent des objets selon certains paramètres. Puis d'appliquer des méthodes analytiques, principalement d'analyse complexe, et probabilistes pour obtenir des renseignements sur les séries génératrices (typiquement le comportement asymptotique des coefficients) et donc sur le problème initial.

La théorie de **Mellin**[7] est un des outils favoris de la combinatoire analytique. Elle permet de faire le passage entre des informations sur le comportement asymptotique d'une fonction f et des informations sur les singularités de la transformée de Mellin $f^\star$ de la fonction f. Pour une présentation plus complète des propriétés de la transformée de Mellin, l'ouvrage de Flajolet, Gourdon et Dumas [FGD95] est une référence. Le livre de Szpankowski [Szp01] y consacre un chapitre complet avec de nombreuses applications. Nous détaillons une application de la transformée de Mellin à la détermination du comportement asymptotique de la longueur de cheminement dans un trie dans la section 1.5.

➤ Résultats.

➽ Trois paramètres principaux des arbres des suffixes construits sur les n premiers suffixes d'un texte aléatoire sont étudiés dans cette thèse : la taille S_n, la longueur de cheminement L_n et la profondeur typique D_n. Nous obtenons le comportement asymptotique de l'espérance et de la variance de ces paramètres. La moyenne de la profondeur typique et de la longueur de cheminement sont liées par la relation

$$n\mathbb{E}(D_n) = \mathbb{E}(L_n). \tag{1}$$

Nous montrons que les moyennes de ces paramètres se comportent au premier ordre asymptotique comme celles de leurs analogues dans les tries. La spécificité des arbres des suffixes n'influe que sur les termes sous-linéaire du comportement asymptotique pour la taille et la longueur de cheminement et que sur des termes en $o(1)$ pour la profondeur typique.

Peu de résultats ont été obtenus sur la taille et la profondeur des arbres des suffixes jusqu'actuellement. Blumer, Ehrenfeucht et Haussler [BEH89] montrent que sous un modèle de source symétrique la taille[8] moyenne est asymptotiquement linéaire avec une faible fluctuation. Jacquet et Szpankowski [JS94] obtiennent le comportement asymptotique de la moyenne de la taille et de la profondeur typique sous un modèle de source sans mémoire biaisé. Ils trouvent aussi la variance et la distribution limite de la profondeur typique. Pour arriver à leur résultats, les auteurs introduisent une nouvelle méthode de *string ruler*. Leur méthode semble difficile à généraliser à des modèles de sources plus généraux.

> Notre but dans le chapitre 2 est donc à la fois
> – de mettre en place une méthodologie susceptible de passer à des modèles de sources plus larges
> – d'améliorer les résultats antérieurs sur les moyennes de la taille, de la longueur de cheminement, de la profondeur typique ainsi que le contrôle du terme d'erreur.
> Dans le chapitre 5, le résultat sur la profondeur typique est étendu au modèle markovien.

Notre approche pour déterminer le comportement asymptotique de la moyenne de la taille, de la profondeur typique et de la longueur de cheminement se base sur la remarque de Jacquet et

[7] Hjalmar Mellin, mathématicien finnois (1854–1933).
[8] Pour Blumer, Ehrenfeucht et Haussler la taille est le nombre de nœuds alors qu'ailleurs la taille est le nombre de nœuds internes

Szpankowski qui énoncent “suffix trees do not differ too much from independent tries. But, tries have been analyzed extensively over last few years, and virtually we know almost everything about them”.

Les trois étapes principales de notre méthode dans chacun des chapitres 2 et 5 sont :

1. donner une expression asymptotique de la moyenne de ces paramètres dans les arbres des suffixes en utilisant les probabilités qu'un texte aléatoire de taille n contienne exactement 0 ou 1 occurrence d'un motif donné w ;
2. estimer l'asymptotique de la différence entre cette expression pour un arbre des suffixes et celle pour un trie en découpant la somme sur tous les motifs ;
3. conclure en se servant des résultats sur les tries.

Dans le chapitre 2 tiré de [Fay04] nous obtenons les comportements asymptotiques de la taille et de la longueur de cheminement sous un modèle de source sans mémoire

$$\boxed{\begin{aligned}\mathbb{E}(S_n) &= \frac{n}{h} + n\epsilon_1(n) + o(n) \text{ et}\\ \mathbb{E}(L_n) &= \frac{n\log n}{h} + Kn + n\epsilon_2(n) + o(n),\end{aligned}} \tag{2}$$

où K est une constante explicitement déterminée, h est l'entropie de la source, les ϵ_i sont des fonctions de très petit module oscillant autour de zéro. Le chapitre 5 issu de [FW05] donne le comportement asymptotique de la profondeur typique sous un modèle de source markovienne d'ordre 1 :

$$\boxed{\mathbb{E}(D_n) = \frac{\log n}{h} + C + \epsilon_3(n) + O(n^{-c}),} \tag{3}$$

où C est une constante explicitement déterminée, $\epsilon_3(n)$ une fonction de très petit module oscillant autour de zéro et c un réel positif. Dans le cas sans mémoire, les constantes C et K sont identiques.

La détermination de la distribution du nombre d'occurrences du motif w dans un texte dont la longueur tend vers l'infini est un domaine de recherche actif. La résolution de cette question sert à obtenir des critères sur le degré d'aléa dans l'apparition d'un motif. Typiquement la question suivante intervient en bio-informatique : la présence de j occurrences du motif w dans le génome est-elle « normale » (*i.e.* conforme au modèle aléatoire) ou possède-t-elle une signification biologique ? Cette question est à l'origine de plusieurs articles dont [FPS96, RS98, NSF99, Sch00, RD04, BCRV05].

Dans le chapitre 3, nous menons l'étude du comportement asymptotique de la variance de la taille et de la longueur de cheminement dans un trie sous un modèle de source sans mémoire. Le comportement asymptotique dans le cas symétrique a déjà été obtenu pour la longueur de cheminement par Kirschenhofer, Prodinger et Szpankowski [KPS89] et pour la taille par Kirschenhofer et Prodinger [KP91]. La moyenne de la profondeur typique D_n et de la longueur de cheminement L_n sont reliées par l'équation (1) ; il n'en est rien pour les variances de ces deux paramètres. Le résultat de [JS94] sur la variance de la profondeur typique ne peut donc servir pour obtenir la variance de la longueur de cheminement. Jacquet et Régnier [JR87, RJ89] ont

montré que la variance de la taille d'un trie est linéaire pour un modèle dans lequel le nombre de texte suit une loi de Poisson de paramètre z et où les textes sont produits par une source sans mémoire biaisée. De plus, ils ont trouvé une expression du coefficient de linéarité.

D'abord nous utilisons une décomposition combinatoire du problème pour montrer qu'asymptotiquement la variance de la taille sous un modèle de Poisson de paramètre z et pour une source sans mémoire se comporte en

$$\boxed{\mathbb{V}_{\mathcal{P}(z)}(S_n) = O(z).} \tag{4}$$

Nous obtenons aussi une première démonstration complète du comportement asymptotique de la longueur de cheminement d'un trie pour une source apériodique sans mémoire biaisée et sous chacun des deux modèles de Poisson de paramètre z et de Bernoulli de paramètre n.

$$\boxed{\begin{aligned} \mathbb{V}_{\mathcal{P}(z)}(L_n) &= \frac{1}{h^2} z \log^2 z + K z \log z + O(z) \quad \text{et} \\ \mathbb{V}_n(L_n) &= \left(\frac{h_2 - h^2}{h^3}\right) n \log n + O(n), \end{aligned}} \tag{5}$$

où la constante K est explicitement déterminée (cf. théorème 6 page 64) et dépend des probabilités p et q.

Ensuite nous obtenons, par une méthode similaire à celle de Guibas et Odlyzko [GO81b], les quatre séries génératrices qui comptent le nombre de textes de taille donnée avec 0 ou 1 occurrence d'un motif w et du motif w'. Les formulations indigestes de ces séries génératrices, et plus particulièrement de leur dénominateur, rend l'extraction des singularités, premier pas vers la détermination du comportement asymptotique de la variance de la taille et de la longueur de cheminement dans un arbre des suffixes, difficile.

➽ Le principe de l'algorithme de compression sans perte décrit précédemment consiste en la création d'un dictionnaire des mots déjà rencontrés lors du parcours du texte pour que, si ces mots reviennent plus ou moins fréquemment, ils puissent être codés par un petit entier. Crochemore, Mignosi, Restivo et Salemi [CMRS99, CMRS00] introduisent un nouveau paradigme d'algorithme de compression de données sans perte appelé `DCA` pour *Data Compression using Antidictionaries*. Cet algorithme se base sur la construction d'un *anti-dictionnaire* qui, au lieu de garder en mémoire les mots déjà vus, stocke un ensemble de mots absents du texte.

Les performances de compression et de décompression rapides (linéaires) rendent l'algorithme `DCA` très attractif. Crochemore[9] *et alii* montrent que, pour certaines sources, l'algorithme `DCA` atteint asymptotiquement le taux entropique de compression h. Crochemore[10] et Navarro [CN02] apportent plus de souplesse à la définition de l'anti-dictionnaire ce qui permet de n'utiliser que 30 à 55 % de l'espace mémoire de l'algorithme « classique. » Ota et Morita [OM04] utilisent le principe de l'anti-dictionnaire pour comprimer sans perte le résultat d'un électrocardiogramme. Ils obtiennent un taux de compression amélioré de 10 % par rapport à un algorithme de type Lempel-Ziv (*i.e.* basé sur un dictionnaire).

La notion fondamentale est celle de mot minimal interdit (*minimal forbidden word* en anglais, ou ici MMI) d'un texte T. Un mot w de taille k est un mot minimal interdit pour le texte T s'il

[9] Rhoooo c'est un complot ! ore␣

[10] Et je ne vous parle pas des algèbres d'Ore !

n'apparaît pas dans le texte T alors que son préfixe et son suffixe de taille $k-1$ s'y trouvent. L'anti-dictionnaire est simplement l'ensemble des MMIs du texte T. Les apparitions du suffixe et du préfixe de taille $k-1$ de w pouvant se chevaucher, l'auto-corrélation va encore[11] jouer un rôle important dans ce problème.

Pour permettre le décodage, l'anti-dictionnaire doit être envoyé avec le code. La taille de l'anti-dictionnaire, c'est-à-dire le nombre de mots dans l'anti-dictionnaire, est donc particulièrement importante dans l'efficacité de l'algorithme. Morita et Ota [MO04] ont montré que la taille de l'anti-dictionnaire, est toujours plus petite que la taille du dictionnaire associé au même texte.

> Dans le chapitre 4, j'obtiens le comportement asymptotique de la moyenne $\mathbb{E}_n(\mathcal{S})$ de la taille d'un anti-dictionnaire construit sur un texte de taille n engendré par une source sans mémoire :
>
> $$\boxed{\mathbb{E}_n(\mathcal{S}) = K\frac{n}{h} + \frac{n}{h}\epsilon(n) + o(n),} \tag{6}$$
>
> où K est une constante explicitement déterminée dépendant des probabilités p et q, h est l'entropie de la source et $\epsilon(n)$ est une fonction oscillant autour de zéro et de très faible amplitude. D'autre part, le modèle approché qui est introduit pour obtenir ce résultat est en lui-même intéressant puisqu'il doit pouvoir s'adapter à d'autres paramètres des anti-dictionnaires.

Ce travail sur la taille de l'anti-dictionnaire fait l'objet d'un article [FFMO06] en préparation avec Philippe Flajolet, Hiroyoshi Morita et Takahiro Ota.

Notes :
- Les fins de preuves sont notées par ◀.
- Le symbole $\sim$ est utilisé pour le terme dominant d'une expression.
- Le symbole $\simeq$ signifie que les deux membres sont « à peu près » égaux.
- Le symbole # désigne le cardinal d'un ensemble.

Bibliographie

[Fay04] Julien Fayolle. An average-case analysis of basic parameters of the suffix tree. In Michael Drmota, Philippe Flajolet, Danièle Gardy, and Bernhard Gittenberger, editors, *Mathematics and Computer Science*, pages 217–227. Birkhäuser, 2004. Proceedings of a colloquium organized by TU Wien, Vienna, Austria, September 2004.

[FW05] Julien Fayolle and Mark Daniel Ward. Analysis of the average depth in a suffix tree under a markov model. In *Proceedings of the 2005 International Conference on the Analysis of Algorithms*, pages 95–104. DMTCS, 2005. Proceedings of a colloquium organized by Universitat Politècnica de Catalunya, Barcelona, Catalunya, June 2005.

[FFMO06] Julien Fayolle, Philippe Flajolet, Hiroyoshi Morita, and Takahiro Ota. Average size of the antidictionary in DCA. *to appear*, 2006.

[11] Ce ore␣ brille comme un sémaph...

Chapitre 1

Objets, modèles et outils

Nous présentons dans ce chapitre les outils et les modèles qui sont utiles tout au long de cette thèse. Le *trie* est une structure de données efficace pour la recherche de chaînes dans un dictionnaire et la mise à jour du dictionnaire. Les paramètres principaux du trie sont définis ainsi que leur signification. L'intérêt de leur étude est mis en avant. Les chaînes sont engendrées par un modèle de source. L'*arbre des suffixes* ou *suffix tree* est une structure de données liée au trie qui sera largement étudiée dans cette thèse. Quelque-unes des *myriades* d'applications des arbres des suffixes sont détaillées dans la section 1.3 comme, par exemple, un algorithme de compression de données sans perte dû à Lempel et Ziv et appelé LZ'77. Je rappelle les comportements asymptotiques de la taille et de la longueur de cheminement dans un trie sous le modèle étudié ici. Ces résultats sont connus depuis longtemps mais nous reviendrons sur les méthodes qui permettent de les obtenir (comme la précieuse *transformée de Mellin*) car elles sont essentielles à cette thèse.

We recall the tools and models that are useful in this thesis. The *trie* is a data-structure highly efficient for querying for strings in a dictionary and updating the dictionary. The main parameters are defined along with their actual meaning. The strings are generated by a source model. My thesis focuses on a data-structure derived from the trie : the *suffix tree*. Some of the *myriad* of applications of a suffix tree, like for instance the LZ'77 lossless data-compression algorithm due to Lempel and Ziv, are described in section 1.3. Then the behaviors of the average path length and size under our model are shown. Although these behaviors are known for some time, we will need them later and it also serves as a reminder for the use of the invaluable *Mellin transform*

Sommaire

1.1 Introduction

La structure d'*arbre digital* ou de *trie* est une structure de donnée qui permet de gérer efficacement la recherche d'un motif dans un ensemble de mots. Elle a été introduite par de la Briandais [dlB59] en 1959 et par Fredkin [Fre60] en 1960. Les mots de l'ensemble de départ sont rangés dans les feuilles d'un arbre suivant leurs lettres successives. Un mot est recherché dans cette structure en le lisant lettre après lettre, comme on le ferait pour rechercher la définition

d'un mot dans un dictionnaire. Les tries ont été largement étudiés depuis Knuth [Knu73]. Le chapitre sur les tries du livre de Mahmoud [Mah92] rend compte des résultats obtenus dans l'étude des tries. La thèse de Clément [Clé00] propose un historique exhaustif et plus récent.

C'est sur un type de structure de donnée lié au trie, l'*arbre des suffixes* ou *suffix tree*, que se basent certains algorithmes essentiels pour l'informatique comme par exemple l'algorithme de compression de données LZ'77 du à Lempel et Ziv [ZL77]. Cette structure d'arbre des suffixes joue donc une grand rôle dans la compression de données qui est l'objet central de cette thèse. La connaissance du comportement des paramètres d'un arbre des suffixes permet de mieux évaluer les performances (de création, d'insertion, de stockage en mémoire) de la structure d'arbre des suffixes et la complexité des algorithmes qui utilisent cette structure de donnée. La principale difficulté de l'analyse des arbres des suffixes tient en la *corrélation* d'un motif : la présence avérée d'un motif à un instant va renseigner sur la possible apparition ultérieure de ce même motif.

On commence par définir les notions qui jalonneront l'étude menée, ainsi que les paramètres pertinents pour l'étude d'un arbre. Le modèle de génération des symboles appelé *source* est développé dans la section 1.2.4. Il existe deux modèles probabilistes sur le nombre d'éléments rangés dans le dictionnaire : le modèle de Bernoulli où le nombre d'éléments est fixe et le modèle de Poisson qui énonce que le nombre d'éléments suit une loi de Poisson. Ce dernier sera détaillé dans la section 1.2.5.

Nous montrons ensuite dans la section 1.3 quelques applications des arbres des suffixes, en particulier en compression de données et en algorithmique du texte. La plus connue de ces applications est probabalement l'algorithme LZ'77 qui crée des fichiers `.zip` ou `.gzip`.

Les résultats sur le comportement asymptotique de la moyenne de la taille et de la longueur de cheminement des tries sous un modèle simple (sans mémoire) sont connus depuis longtemps (Knuth [Knu73] et Flajolet [Fla83] pour le cas symétrique et Jacquet et Régnier [JR86, RJ89] pour le cas biaisé). Le terme dominant du comportement asymptotique de la taille est n/h sur un ensemble de n mots et $(1/h)n\log n$ pour le longueur de cheminement pour une constante h dépendante de la source. Néanmoins il m'a semblé intéressant (dans les sections 1.4 et 1.5) de revenir sur l'analyse de ces deux paramètres pour montrer les méthodes employées. Ces méthodes servent d'ailleurs dans les chapitre ultérieurs. La théorie de Mellin par exemple est utilisée un grand nombre de fois dans cette thèse.

1.2 Objets et modèles

1.2.1 Notions de base sur les tries

Définition 1 *Les* tries[12] *ou* arbres digitaux *sont définis récursivement sur un ensemble de mots X de l'alphabet $\mathcal{A} = \{a_1, a_2, ..., a_m\}$ par :*

$$\text{trie}(X) = \begin{cases} \emptyset & \text{si } |X| = 0, \\ \bullet & \text{si } |X| = 1, \\ \langle \bullet, \text{trie}(X\backslash a_1), ..., \text{trie}(X\backslash a_m)\rangle & \text{sinon.} \end{cases}$$

L'ensemble $X\backslash\alpha$ est le sous-ensemble des mots de X commençant par la lettre α auxquels on retire leur lettre initiale α.

[12] Ce mot est un mélange entre *tree* (arbre) et *retrieve* (récupérer) car le trie est une structure de donnée efficace pour la gestion d'un dictionnaire, et en particulier pour retrouver un mot.

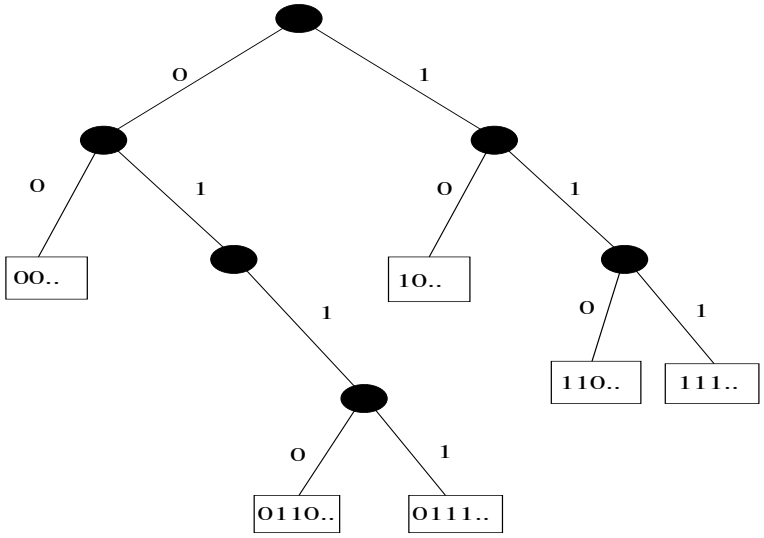


FIG. 1.1 – Exemple de trie sur 2 lettres. Les branches du trie sont étiquetées par les lettres de l'alphabet.

Remarque : Pour éviter le cas où un mot de X est préfixe d'un autre mot, nous ne regardons que des mots infinis. Les mots X sont ainsi stockés dans les feuilles (nœuds sans fils) du trie. Dans la pratique, les mots sont évidemment finis. Pour éviter ce problème de mot préfixe d'un autre, une lettre artificielle qui n'appartient pas à l'alphabet de base $\mathcal{A}$, typiquement notée \$ ou #, est concaténée à la fin de chaque mot de l'ensemble X.

Définition 2 *La longueur $|w|$ du mot w est le nombre de symboles du mot.*

Dans le reste de cette thèse, l'alphabet $\mathcal{A}$ sur lequel les textes sont construits est binaire.

Le trie de la figure 1.1 est construit sur l'alphabet binaire $\mathcal{A} = \{0, 1\}$ et sur un ensemble X de six mots infinis

$$X = \{0010001..., 0110101..., 0111001..., 10010010..., 110000..., 1110010...\}.$$

L'ensemble X est scindé en deux : d'un côté les mots dont la première lettre est 0, soit

$$A = \{0010001..., 0110101..., 0111001...\}$$

et de l'autre ceux dont la première lettre est 1, soit

$$B = \{10010010..., 110000..., 1110010...\}.$$

Dans cette première étape de la construction récursive, l'arbre a une racine et, puisqu'aucun des deux sous-ensembles A et B n'est vide, deux sous-tries. Le sous-trie gauche de l'arbre est

$$\text{trie}(A\backslash 0) = \text{trie}(\{010001..., 110101..., 111001...\})$$

et le sous-trie droit

$$\text{trie}(B\backslash 1) = \text{trie}(\{0010010..., 10000..., 110010...\}).$$

Le procédé de séparation est itéré tant que les ensembles considérés ne sont pas réduits à un seul mot ou vide. À la fin du processus, on obtient le trie de la figure 1.1. On remarque qu'aucun mot de l'ensemble X ne commence par le préfixe `010` et cela se retrouve dans le trie : le sous-trie gauche du nœud placé à la « position » `01` est vide.

D'après la définition, le sous-trie trie($X\backslash\alpha$) est construit sur l'ensemble des mots de X qui commencent par la lettre α auxquels on retire leur lettre initiale. La définition peut-être itérée, le sous-trie trie($X\backslash\alpha\beta$) est constitué des mots de X commençant par la lettre α suivie de la lettre β, c'est-à-dire des mots commençant par le préfixe $\alpha\beta$.

Soit O un nœud interne du trie à profondeur k. Il existe un unique mot de taille k qui permet d'accéder au nœud O en lisant successivement les étiquettes des branches du trie. Ce mot est noté $w(O)$, il est le préfixe commun à tous les mots de X dont la feuille associée (la feuille du trie qui contient le mot) est dans le sous-trie de racine O. La position O dans l'arbre binaire infini complet est un nœud interne du trie si au moins deux chaînes de l'ensemble de base X commencent par la préfixe $w(O)$.

1.2.2 Les paramètres du trie

Dans cette section, les expressions de certains paramètres classiques des arbres sont données pour un trie. À cet effet, le paramètre N_w est introduit. Il est défini sur l'ensemble des mots qui forment le trie. Soit X un ensemble de n textes infinis, $N_w(X)$ vaut le nombre de textes de X qui commencent par w. Le paramètre N_w va permettre de décrire élégamment la plupart des paramètres du trie. Par simplicité d'écriture et quand l'ensemble de base sera évident, on écrira juste N_w au lieu de $N_w(X)$.

On rappelle la définition

Définition 3 *Soit* **P** *une propriété, les crochets d'Iverson sont définis par :*

$$[\![\mathbf{P}]\!] = \begin{cases} 1 & \textit{si } \mathbf{P} \textit{ est vraie,} \\ 0 & \textit{sinon .} \end{cases}$$

Taille : La taille S d'un trie est définie comme le nombre de nœuds internes dans le trie formé sur l'ensemble de mots X. Cette variable aléatoire est liée à la place que prend le stockage en mémoire d'un trie. Il y a un nœud interne dans le trie à la position O (position définie par rapport à un arbre binaire complet infini), si au moins deux mots de l'ensemble X commencent par $w = w(O)$, c'est-à-dire $N_w(X) \geq 2$, ainsi :

$$S = \sum_{w \in \mathcal{A}^\star} [\![N_w \geq 2]\!]. \tag{1.1}$$

Longueur de cheminement externe : La longueur de cheminement, notée L, est définie comme la somme sur chaque feuille de l'arbre de la distance de la racine à la feuille. Elle est reliée au temps de construction du trie et au temps moyen de recherche d'une chaîne avec succès.

La distance de la racine à une feuille est le nombre de nœuds internes situés sur le trajet de la racine à la feuille. Un nœud interne O est compté dans la longueur de cheminement autant de fois qu'il y a de mots commençant par $w = w(O)$, soit $N_w(X)$ fois. Ceci est vrai pour chaque nœud interne. On obtient :

$$L = \sum_{w \in \mathcal{A}^\star} N_w [\![N_w \geq 2]\!]. \tag{1.2}$$

Hauteur : La hauteur H est la plus grande longueur d'un chemin allant de la racine à une feuille du trie. La hauteur est le temps maximum mis pour la recherche avec succès d'un mot dans le dictionnaire (implanté par un trie). Il n'y a pas nécessairement une unique feuille à hauteur H dans le trie. La hauteur est aussi la longueur du préfixe qui amène à un nœud le plsu éloigné de la racine plus un (pour arriver à la feuille), d'où :

$$H = \max\{|w| \ : \ N_w \geq 2\} + 1. \tag{1.3}$$

Nombre de nœuds internes à profondeur k : Le profil de l'arbre, c'est-à-dire le nombre de nœuds à profondeur donnée est un moyen d'unifier l'étude des paramètres précédents. Le profil permet aussi de connaître la forme de l'arbre.

Il existe un nœud interne O à profondeur k si au moins deux mots de X commencent par le préfixe $w(O)$, soit $N_{w(O)} \geq 2$. On en infère :

$$N(k) = \#\{w \in \mathcal{A}^k : N_w \geq 2\}. \tag{1.4}$$

Niveau de saturation : C'est la profondeur maximale à laquelle le trie est complet. Le niveau f est dit de saturation s'il est saturé (ou complet) mais pas le niveau suivant $(f+1)$. L'introduction de ce paramètre est justifiée par des économies de mémoire : si on connait le niveau de saturation (*fill-up level* en anglais) alors on peut remplacer les nœuds internes du trie de profondeur plus petite que f qui, par définition du niveau de saturation, existent tous, par un tableau contenant les nœuds du niveau f et en insérant les nouvelles chaînes à partir des nœuds de ce tableau (c'est le principe, appliqué récursivement, du *level compressed trie* ou LC-trie proposé et analysé par Andersson et Nilsson [AN93] et analysé plus finement par Devroye [Dev01]).

Pour exprimer ce paramètre, on remarque que le niveau f de saturation doit comporter au moins un nœud n'ayant pas deux fils. Ce nœud peut être une feuille (si ses deux fils sont vides) ou un nœud interne avec un seul fils vide. Dire que le nœud O associé au préfixe w n'a pas deux fils dans l'arbre signifie que w est préfixe d'au moins une chaîne de X et que soit aucune chaîne de X ne commence par $w\mathtt{0}$, soit aucune chaîne de X ne commence par $w\mathtt{1}$, d'où :

$$f = \min\{|w| \ : \ N_{w\mathtt{0}} = 0 \text{ ou } N_{w\mathtt{1}} = 0\}. \tag{1.5}$$

1.2.3 Arbre des suffixes

Les arbres des suffixes se construisent sur le même schéma récursif de discrimination selon les premières lettres que les tries (cf. définition 1), mais l'ensemble qui sert de base à la construction du trie suffixe (nommé X pour le trie) est bâti autrement. Les paramètres du trie de la section 1.2.2 sont aussi intéressants pour un arbre des suffixes.

Un trie est construit sur un ensemble X de mots infinis sur l'alphabet $\mathcal{A}$. Pour un arbre des suffixes les choses sont plus délicates : on se donne un texte infini T sur l'alphabet $\mathcal{A}$ et un entier n. L'ensemble Y_n est l'ensemble des n premiers suffixes du texte T. Un arbre des suffixes est construit sur la même règle récursive que le trie à partir de l'ensemble Y_n qui compte n chaînes infinies (donc la définition 1 s'applique bien). Le k^{e} suffixe infini du texte T est le mot infini qui commence à partir de la position k du texte T.

Exemple : Si $T = \mathtt{01010111010011}\ldots$ et $n = 6$ alors les 6 premiers suffixes de la chaîne T sont : le texte T lui-même, premier suffixe trivial du texte T et les suffixes successifs

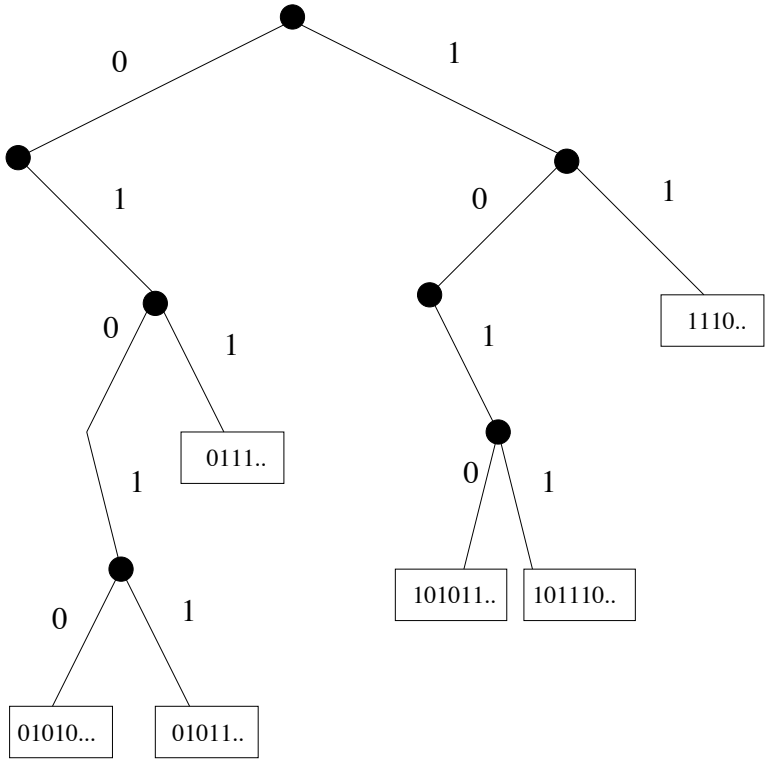


FIG. 1.2 – Exemple de d'arbre des suffixes construit sur les 6 premiers suffixes de $T =$ `01010111010011...`. Les branches du trie sont étiquetées par les lettres de l'alphabet.

```
1010111010011...,
 010111010011...,
  10111010011...,
   0111010011... et
    111010011...
```

La paramètre $\widehat{N}_w$ est défini sur les couples (T, n) où T est un texte infini construit sur l'alphabet $\mathcal{A}$ et n un entier non nul. L'ensemble Y_n est formé des n premiers suffixes du texte T. La valeur de $\widehat{N}_w(T, n)$ est en fait $N_w(Y_n)$, soit le nombre de mots de Y_n qui commencent par w. Le nombre de mots de Y_n qui commencent par w est aussi le nombre de suffixes de T qui commencent par w, c'est-à-dire le nombre d'occurrences du motif w dans les n premières positions du texte T. Comme pour le paramètre N_w, les arguments T et n ne sont pas mentionnés dans l'écriture de $\widehat{N}_w$ pour que le texte reste digeste. À partir de cette analogie entre N_w et $\widehat{N}_w$, on obtient pour le trie et l'arbre des suffixes les expressions suivantes :

Taille :

$$\sum_{w\in\mathcal{A}^\star} [\![N_w \geq 2]\!] \text{ et } \sum_{w\in\mathcal{A}^\star} [\![\widehat{N}_w \geq 2]\!].$$

Longueur de cheminement :

$$\sum_{w\in\mathcal{A}^\star} N_w[\![N_w \geq 2]\!] \text{ et } \sum_{w\in\mathcal{A}^\star} \widehat{N}_w[\![\widehat{N}_w \geq 2]\!].$$

Hauteur :

$$\max\{|w| \ : \ N_w \geq 2\} + 1 \text{ et } \max\{|w| \ : \ \widehat{N}_w \geq 2\} + 1.$$

Nombre de nœuds internes à profondeur k :

$$\#\{w : |w| = k \text{ et } N_w \geq 2\} \text{ et } \#\{w : |w| = k \text{ et } \widehat{N}_w \geq 2\}.$$

Niveau de saturation :

$$\min\{|w| \ : \ N_w = 1\} \text{ et } \min\{|w| \ : \ \widehat{N}_w = 1\}.$$

Exemple : Pour les deux exemples précédents (le trie de la figure 1.1 et l'arbre des suffixes de la figure 1.2) nous donnons les valeurs des cinq paramètres qui viennent d'être introduits.

- La taille du trie est 6 et celle de l'arbre des suffixes 8.
- La longueur de cheminement du trie est 18 et celle de l'arbre des suffixes 23.
- La hauteur du trie vaut 4 et celle de l'arbre des suffixes 5.
- Pour, par exemple, la profondeur 3, le trie a 1 nœud interne à profondeur 3 et l'arbre des suffixes 2.
- Le niveau de saturation du trie est 1 et celui de l'arbre des suffixes 1.

1.2.4 Sources

Les paramètres définis dans les sections précédentes le sont à partir d'un ou de plusieurs textes. Comment obtient-on ces textes infinis ? À l'aide d'une *source*, c'est-à-dire d'un procédé qui produit des lettres de l'alphabet de base $\mathcal{A}$ de manière aléatoire. L'émission des symboles dépend des caractéristiques de la source. Dans toute cette thèse, l'alphabet de base est binaire. Il est noté $\mathcal{A} = \{\mathtt{0}, \mathtt{1}\}$. Ce choix permet de simplifier à la fois les calculs et les notations. La nature des résultats exposés ne s'en trouve jamais modifiée. Deux types de sources interviennent dans cette thèse : les sources sans mémoire et les sources markoviennes.

Dans le cas d'une source sans mémoire, la distribution du symbole émis au temps t ne dépend pas des lettres émises précédemment. La source sans mémoire (p, q) produit le symbole `0` avec la probabilité p et le symbole `1` avec la probabilité q (dans la suite, on suppose que p est supérieur à q). Si $p = q = \frac{1}{2}$, la source est dite symétrique. Elle est dite biaisée sinon. En termes plus probabilistes, le modèle de source sans mémoire (p, q) signifie que les lettres émises par la source sont déterminées par des variables aléatoires indépendantes prenant les valeurs `0` ou `1` et suivant toutes une loi de Bernoulli de paramètre p.

Une source est markovienne d'ordre k (k entier) si la distribution du symbole aléatoire émis au temps t dépend des k symboles émis précédemment par la source. Le cas le plus simple est la source de Markov d'ordre 1 où le symbole aléatoire émis dépend uniquement du symbole qui le précède (et de l'aléa). L'ordre k de la source est le nombre de lettres qui entrent en jeu dans la distribution du symbole qui va être émis. Les lettres produites par la source peuvent aussi être vues comme les valeurs de variables aléatoires indépendantes prenant les valeurs `0` et `1` et suivant toutes une même loi de Markov d'ordre k. Les sources markoviennes seront utilisées uniquement dans le chapitre 6. Nous considérons que le lecteur connait les définitions de la matrice de transition, du vecteur stationnaire et autres définitions de base des chaînes de Markov (des milliers d'ouvrages le font très bien, la référence la plus classique est probablement le livre de William Feller [Fel68]).

Définition 4 *Soient w un mot (éventuellement infini) sur l'alphabet $\mathcal{A}$ et une source $\mathcal{S}$, p_w est défini comme la probabilité qu'un mot émis par la source commence par le préfixe w. La probabilité p_w se nomme probabilité d'occurrence associée au préfixe w.*

Exemples : Dans le cas d'une source sans mémoire symétrique, la probabilité d'occurrence du motif w=010111110101 se calcule en prenant le 0 initial avec la probabilité $\frac{1}{2}$, puis le 1 en deuxième position avec la probabilité $\frac{1}{2}$, etc... Au final, on obtient une probabilité $2^{-|w|} = 2^{-12}$. Sur cette source particulière, la probabilité d'occurrence du motif va être indépendante des symboles composant le motif (seule la longueur du motif entre en compte).

Autre exemple : le cas d'une source sans mémoire biaisée (p, q). Avec le même motif w, la probabilité d'occurrence est p^4q^8. Il s'agit pour la source de choisir le premier symbole qui est 0 et qui apparaît avec la probabilité p, le deuxième symbole est 1 qui apparaît avec la probabilité q et ainsi pour les autres symboles.

Troisième exemple et toujours avec le même motif w : une source markovienne d'ordre 1. Le vecteur de probabilité stationnaire π et la matrice de transition P valent respectivement

$$\pi = \begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix} = \begin{pmatrix} \frac{3}{17} \\ \frac{14}{17} \end{pmatrix} \text{ et } P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 0.3 & 0.15 \\ 0.7 & 0.85 \end{pmatrix}.$$

De plus, la source a atteint son régime stationnaire. Dans ce modèle, la probabilité d'occurrence du motif w s'écrit

$$\pi_0 p_{01}p_{10}p_{01}p_{11}p_{11}p_{11}p_{11}p_{10}p_{01}p_{10}p_{01} = \pi_0(p_{01})^4(p_{10})^3(p_{11})^4 = \frac{3}{17}(0.15)^4(0.7)^3(0.85)^4.$$

1.2.5 Poissonisation

Les paramètres N_w et $\widehat{N}_w$ sont définis pour un nombre donné de textes. Le modèle probabiliste sur les textes est le modèle de source défini au paragraphe précédent, nous discutons maintenant du modèle probabiliste sur le nombre d'éléments (nombre de textes dans X ou nombre de suffixes de T) dans l'ensemble. Deux modèles sur le nombre d'éléments seront utilisés dans cette thèse : le modèle de Bernoulli et le modèle de Poisson[13]. La variable aléatoire $\mathfrak{N}$, définie sur l'ensemble des entiers, détermine le nombre d'éléments à tirer (ou le nombre de suffixes à regarder).

Dans le modèle de Bernoulli, le nombre d'éléments est fixé et vaut n. La variable aléatoire $\mathfrak{N}$ suit alors une loi de Dirac de paramètre n sur l'ensemble des entiers, façon savante de dire $\mathbb{P}(\mathfrak{N} = n) = 1$ ou $\mathfrak{N} = n$.

Dans le modèle de Poisson, la variable aléatoire $\mathfrak{N}$ suit une loi de Poisson de paramètre z, soit

$$\forall j \in \mathbb{N},\ \mathbb{P}(\mathfrak{N} = j) = \frac{z^j e^{-z}}{j!}.$$

La moyenne du nombre de mots dans un ensemble est z, et puisque l'écart-type est petit devant la moyenne ($\sigma(\mathfrak{N}) = \sqrt{z}$), le nombre de mots émis va rester proche de la valeur z avec forte probabilité.

L'introduction de la loi de Poisson pour compter le nombre d'éléments dans un problème est couramment appelée *poissonisation* du problème. Dans son livre [Szp01], Spankowski consacre un chapitre à la poissonnisation/dépoissonnisation. Il fait remonter l'introduction de la technique de poissonnisation à Marek (Mark) Kac [Kac49] en 1949.

Il y a deux avantages à la poissonisation : la simplification des calculs par rapport au cas Bernoulli et l'utilisation d'une variable z qui autorise le recours aux méthodes de la combinatoire analytique. La simplification dans le cas Poisson tient au fait que la loi de Poisson se transmet aux

[13] Siméon Denis Poisson, mathématicien, géomètre et physicien français (1781–1840)

variables aléatoires N_w avec une modification simple du paramètre. De plus il y a indépendance de deux variables aléatoires entre deux motifs ayant la même longueur :

Proposition 1 *Si le nombre d'éléments de l'ensemble X suit une loi de Poisson de paramètre z notée $\mathcal{P}(z)$ alors la variable aléatoire N_w suit une loi de Poisson de paramètre zp_w. De plus, pour deux préfixes w et w' de même longueur, les variables aléatoires N_w et $N_{w'}$ sont indépendantes.*

Preuve : La variable aléatoire $\mathfrak{N}$ suit une loi de Poisson de paramètre z. Pour k et n deux entiers avec $k \leq n$, on a :

$$\mathbb{P}(N_w = k | \mathfrak{N} = n) = \binom{n}{k} p_w^k (1 - p_w)^{n-k}.$$

En effet parmi les n chaînes de l'ensemble X, on en choisit k qui vont avoir le préfixe w, les $n - k$ autres ne commençant pas par ce préfixe. La probabilité de l'événement $\{N_w = k\}$ est la somme des probabilités conditionnelles :

$$\begin{aligned}\mathbb{P}(N_w = k) &= \sum_{n \geq k} \mathbb{P}(\mathfrak{N} = n) \binom{n}{k} p_w^k (1 - p_w)^{n-k} = \sum_{n \geq k} \frac{z^n e^{-z}}{n!} \binom{n}{k} p_w^k (1 - p_w)^{n-k} \\ &= \frac{(zp_w)^k}{k!} e^{-p_w z}.\end{aligned}$$

Donc la variable aléatoire N_w suit une loi de Poisson de paramètre $p_w z$. L'indépendance des variables aléatoires N_w et $N_{w'}$ se montre de manière analogue. ◄

Dans la suite, pour bien différencier les modèles probabilistes sur le nombre de chaînes que l'on regarde, l'espérance sous modèle de Poisson de paramètre z sera notée $\mathbb{E}_{\mathcal{P}(z)}$ et l'espérance sous modèle de Bernoulli avec n chaînes $\mathbb{E}_n$. Ces notations seront aussi valables pour les probabilités. S'il n'y a pas de confusion possible ou quand la nature du modèle n'entre pas en jeu, ces indices seront omis.

1.3 Applications

Les arbres des suffixes ont été introduits par Weiner [Wei73] en 1973 pour accélerer les opérations de recherche de motifs dans un texte. Dans cette section, on met en avant quelque-unes des applications des arbres des suffixes à l'informatique. La plus importante est probablement l'algorithme de compression sans perte (appelé aussi conservative) de Ziv et Lempel en 1977 [ZL77]. L'algorithme est couramment appelé LZ'77. L'article d'Apostolico [Apo85] fournit de nombreuses applications de l'arbre des suffixes. Le livre de Gusfield [Gus97] est assez récent et présente de manière pédagogique des applications des arbres des suffixes.

Recherche d'un mot dans un texte :

Soit T un texte et w un mot. La recherche du mot w dans le texte peut s'effectuer efficacement à l'aide d'un arbre des suffixes. Une fois l'arbre des suffixes du texte T construit, la recherche du mot w est juste une descente le long des branches de l'arbre des suffixes en lisant les lettres successives de w (comme dans un dictionnaire papier). Si la descente de toutes les lettres de w nous amène sur un nœud de l'arbre (feuille ou nœud interne), c'est que le mot est présent dans le texte. L'arbre des suffixes peut-être modifié de manière à répondre à d'autres questions : combien de fois le motif w apparaît dans le texte ou encore où sont les occurrences de w dans le texte.

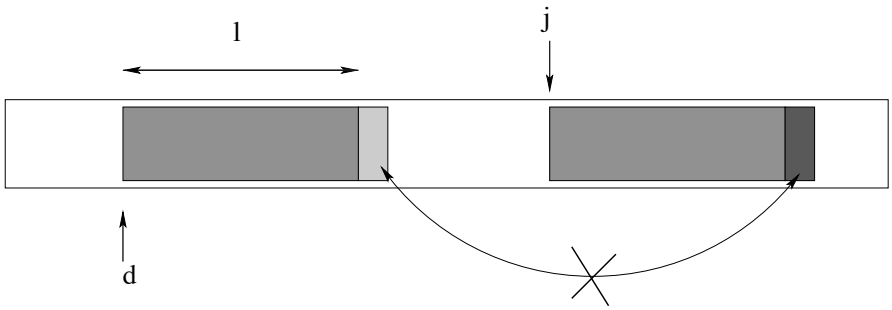


FIG. 1.3 – Principe de l'algorithme LZ'77 : on cherche le plus long mot déjà vu.

Algorithme de compression LZ'77 :

L'algorithme LZ'77 est l'un des algorithmes de compression sans perte les plus utilisés dans la pratique. Il se base sur une structure de données d'arbre des suffixes. L'algorithme `DEFLATE` (RFC 1951) combine les algorithmes de compression LZ'77 et Huffman. Les logiciels `PKZIP`, `WinZip` et `gzip` utilisent `DEFLATE`. Les images au format `Portable Networks Graphics` ou `png` (RFC 2083) sont aussi obtenues par l'algorithme `DEFLATE`. L'implantation de l'algorithme LZ'77 dans ce logiciel n'est pas la version originale de Ziv et Lempel mais une version améliorée appelée Lempel-Ziv-Storer-Szymanski [SS82].

Le manuel de la commande en-ligne `gzip` note :

> Typically, text such as source code or English is reduced by 60-70%. Compression is generally much better than that achieved by LZW (as used in `compress`).

Comment fonctionne l'algorithme LZ'77 ? On se donne un fichier à compresser que l'on appelle *texte*. Le principe de la compression dans cet algorithme est de coder le texte en construisant dynamiquement (en même temps que l'on lit le texte) un *dictionnaire* des mots déjà rencontrés dans le texte. Le texte sera codé par les références aux mots déjà stockés dans le dictionnaire, c'est-à-dire déjà vus dans le texte. Le décompresseur retransforme le code qu'on lui envoie en le texte original (sans perte) en maintenant un dictionnaire en parallèle.

Décrivons (cf. figure 1.3) le fonctionnement de la compression à l'étape i et en ayant traité les j premières lettres du texte ($j \geq i$ et très souvent l'inégalité est stricte). L'arbre des suffixes sur les $j-1$ premiers suffixes du texte déjà lu a été construit lors des étapes antérieures. C'est ce qu'on appelle le dictionnaire. On recherche dans le dictionnaire le plus long mot u correspondant à un facteur commençant en j. La structure d'arbre des suffixes permet une recherche efficace de ce plus long facteur déjà rencontré. Il suffit en effet de descendre dans l'arbre des suffixes selon les lettres du suffixe qui commence en j et de s'arrêter sur le dernier nœud existant dans le chemin. Ce nœud (ce n'est pas nécessairement une feuille) contient des informations comme la position de début de la chaîne u dans le texte (a priori il existe plusieurs occurrences de la chaîne u dans les j premières lettres du texte, cette redondance est utilisée pour corriger d'éventuelles erreurs dans [LS03] et analysée dans [WS04, LSW04, WS05]) La sequence codée envoyée est un triplet constitué de la position du facteur dans le passé du texte (indice de début d et taille l ou indice de début et de fin) et du nouveau symbole (situé après le facteur commençant en j) en position $j+d$. L'ensemble de toutes ces séquences est le texte compressé ou *code*. Le dictionnaire est mis à jour de manière à contenir les $j+d$ premiers suffixes du texte déjà lu (c'est-à-dire de $T_1 \dots T_{j+d}$). Et hop `i++` et `j+=d+1` ! Le curseur se place juste après la nouvelle lettre, soit en position $j+d+1$.

Regardons un exemple simple : le cas où le texte est constitué exclusivement de `0`, la position

du curseur (en j, marqué par $\bullet$) selon l'étape est :

1	$\bullet$0000000000000000000000000000000000	position 1	code $(d = 0, l = 0,$ 0)
2	0$\bullet$0000000000000000000000000000000000	position 2	code (1, 1, 0)
3	000$\bullet$0000000000000000000000000000000000	position 4	(1, 3, 0)
4	0000000$\bullet$0000000000000000000000000000	position 8	(1, 7, 0)
5	0000000000000000$\bullet$0000000000000000000	position 16	(1,15, 0)
6	00000000000000000000000000000000$\bullet$000	position 32	(1, 31, 0)

Deux idées améliorent le rendement pratique de l'algorithme. D'abord, l'utilisation d'une fenêtre coulissante ou *sliding window* de taille fixée : au lieu de chercher le plus long facteur commençant en j qui coïncide avec un mot déjà rencontré dans le texte en entier, on ne regarde qu'un passé de taille finie (disons m) du texte. Le dictionnaire est alors construit sur les m dernières lettres du texte lu $T_{j-m} \dots T_{j-1}$. Cela limite la taille du dictionnaire et par conséquent augmente la vitesse de recherche. L'autre amélioration, due à Storer et Szymanski en 1982, est le traitement des cas où le plus long mot déjà rencontré est réduit à une seule lettre. Leur astuce de programmation consiste juste à utiliser un drapeau pour noter si le *match* est réduit à une seule lettre ou s'il faut donner le triplet (d, l, T_{j+d}) au complet. L'implantation la plus classique de LZSS est due à Bell [Bel86], elle utilise les arbres binaires.

Recherche de *repeats* en bio-informatique :

Les bio-informaticiens et les biologistes sont intéressés à trouver les occurrences d'un même mot (ou un représentant d'un ensemble de mots) dans un ou plusieurs textes (typiquement des génomes). Ce mot peut avoir comme rôle biologique de servir de marqueur d'une portion biologiquement signifiante du génome. Un génome est un texte sur un alphabet de quatre lettres appelées nucléotides ou bases (A, T, C et G). Les génomes de différents êtres vivants sont entrés dans de pantagruéliques bases de données regroupées au sein de l'*International Nucleotide Sequence Database Collaboration* qui comprend les bases de données :

- *GenBank* aux États-Unis `http://www.ncbi.nlm.nih.gov/Genbank/index.html` ;
- *DNA DataBank of Japan* ou DDBJ au Japon `http://www.ddbj.nig.ac.jp/` ;
- *European Molecular Biology Laboratory Data Library* en Europe `http://www.ebi.ac.uk/`.

Depuis août 2005, ces trois bases de données contiennent plus de 100 milliards de paires de bases. La taille de ces bases de données nécessite des algorithmes rapides.

D'après le livre de Waterman [Wat95] (qui est une référence pour la bio-informatique), les arbres des suffixes ont étés introduits en bio-informatique par Martinez [Mar83] en 1983 pour permettre une recherche efficace de *repeats* dans une séquence d'ADN. Les repeats sont des mots qui apparaissent à plusieurs reprises dans le texte.

1.4 Analyse des tries

Dans cette section, l'expression de l'espérance de la taille et de la longueur de cheminement est déterminée sous les modèles de Poisson, de Bernoulli, et avec les sources symétrique ou biaisées (p, q). Sous le modèle de Poisson de paramètre z, l'expression de la moyenne de la taille et de la longueur de cheminement est fonction du paramètre z. La théorie de la transformation de Mellin appliquée à ces fonctions permet de recouvrer leur comportement asymptotique quand z tend vers l'infini. Pour un modèle de Bernoulli de paramètre n, le terme dominant du comportement asymptotique de la taille est en $\frac{n}{h}$ (lemme 2) et celui de la longueur de cheminement $\frac{n \log n}{h}$ (lemme 1), où $h = -p \log p - q \log q$ est l'entropie de la source sans mémoire. Nous n'effectuons

le raisonnement que sur l'espérance de la longueur de cheminement d'un trie. La méthode est aussi valable pour l'espérance de la taille d'un trie.

Il existe deux modèles sur le nombre de textes qui constituent l'ensemble de base X : le modèle de Bernoulli et le modèle de Poisson. La source qui engendre le texte peut être de deux types : symétrique ou biaisée. Dans les quatre cas (modèle sur le nombre de textes et sur le type de la source), on donne l'expression de la moyenne de la longueur de cheminement.

La longueur de cheminement L d'un trie vient d'être définie dans la section 1.2.2 par la formule :

$$L = \sum_{w \in \mathcal{A}^\star} N_w [\![N_w \geq 2]\!],$$

où N_w compte le nombre de mots de l'ensemble de base commençant par le motif w. En se servant de la linéarité de la moyenne, on obtient

$$\begin{aligned} \mathbb{E}(L) = \sum_{w \in \mathcal{A}^\star} \mathbb{E}(N_w [\![N_w \geq 2]\!]) &= \sum_{w \in \mathcal{A}^\star} \mathbb{E}(N_w) - \mathbb{E}(N_w [\![N_w = 1]\!]) - \mathbb{E}(N_w [\![N_w = 0]\!]) \\ &= \sum_{w \in \mathcal{A}^\star} \mathbb{E}(N_w) - \mathbb{P}(N_w = 1), \end{aligned} \tag{1.6}$$

puisque $N_w [\![N_w = 0]\!] = 0$ et $N_w [\![N_w = 1]\!] = [\![N_w = 1]\!]$.

Si on se place dans le cas d'un modèle de Bernoulli avec un nombre n fixé de mots dans l'ensemble X, il faut interpréter la probabilité $\mathbb{P}_n(N_w = 1)$ de manière combinatoire. L'ensemble $\{N_w = 1\}$ est l'ensemble des textes où un seul des n mots produits par la source commence par le préfixe w, c'est le choix d'un mot parmi n, d'où :

$$\mathbb{P}_n(N_w = 1) = \binom{n}{1} p_w (1 - p_w)^{n-1} = n p_w (1 - p_w)^{n-1}. \tag{1.7}$$

Dans le cas d'un modèle de Poisson de paramètre z, la probabilité vaut simplement $z p_w \exp(-z p_w)$. La probabilité qu'exactement k chaînes commencent par le motif w s'exprime comme le nombre de choix de k chaînes commençant par le préfixe w parmi les n fournies par la source alors que les $k - n$ autres textes ne commencent pas par le motif w. La moyenne s'écrit

$$\mathbb{E}_n(N_w) = \sum_{k \geq 0} k \mathbb{P}_n(N_w = k) = \sum_{k=0}^{n} k \binom{n}{k} p_w^k (1 - p_w)^{n-k} = n p_w. \tag{1.8}$$

Sous un modèle de Poisson de paramètre z, la variable aléatoire N_w suit une loi de Poisson de paramètre $z p_w$ et sa moyenne vaut

$$\mathbb{E}_{\mathcal{P}(z)}(N_w) = \sum_{k \geq 0} k \mathbb{P}_{\mathcal{P}(z)}(N_w = k) = \sum_{k \geq 0} \frac{k (z p_w)^k \exp(-z p_w)}{k!} = z p_w. \tag{1.9}$$

Il ne reste plus qu'à particulariser ces formules selon le modèle de source qui nous intéresse :

– Dans le cas d'une source sans mémoire symétrique et d'un modèle de Poisson, les probabilités d'occurrence d'un motif de taille k valent toutes 2^{-k}, donc :

$$\begin{aligned} \mathbb{E}_{\mathcal{P}(z)}(L) &= \sum_{k \geq 0} \sum_{w \in \mathcal{A}^k} \frac{z}{2^k} \left(1 - \exp\left(-2^{-k} z\right)\right) \\ &= \sum_{k \geq 0} z \left(1 - \exp\left(-2^{-k} z\right)\right). \end{aligned} \tag{1.10}$$

– Dans le cas d'une source sans mémoire biaisée (p,q) les $\binom{k}{i}$ préfixes de taille k s'écrivant avec i occurrences du symbole `0` et $(k-i)$ occurrences de `1` ont tous pour probabilité $p^i q^{k-i}$, donc :

$$\begin{aligned}\mathbb{E}_{\mathcal{P}(z)}(L) &= \sum_{k\geq 0}\sum_{w\in\mathcal{A}^k} zp_w\left(1-\exp(-p_w z)\right)\\ &= \sum_{k\geq 0}\sum_{i=0}^{k} z\binom{k}{i}p^i q^{k-i}\left(1-\exp\left(-p^i q^{k-i}z\right)\right).\end{aligned} \tag{1.11}$$

Le même principe de regroupement des motifs de taille k selon leur probabilité est valable pour un trie construit sur un nombre de textes fixé (sous modèle de Bernoulli). Il vient

$$\mathbb{E}_n(L) = \sum_{k\geq 0} n2^k\left(1-\left(1-2^{-k}\right)^{n-1}\right) \text{ et pour une source biaisée,}$$

$$\mathbb{E}_n(L) = \sum_{k\geq 0}\sum_{i=0}^{k} n\binom{k}{i}p^i q^{k-i}\left(1-\left(1-p^i q^{k-i}\right)^{n-1}\right).$$

On récapitule l'expression de la longueur de cheminement sous les différents modèles présentés :

Lemme 1 *Pour une source symétrique et sous modèle de Poisson de paramètre z, la moyenne de la longueur de cheminement L d'un trie a pour expression*

$$\mathbb{E}_{\mathcal{P}(z)}(L) = \sum_{k\geq 0} z\left(1-\exp\left(-2^{-k}z\right)\right) \qquad \textit{et} \qquad \mathbb{E}_n(L) = \sum_{k\geq 0} n\left(1-\left(1-2^{-k}\right)^{n-1}\right)$$

dans un modèle de Bernoulli. Pour une source biaisée (p,q), la moyenne de la longueur de cheminement d'un trie se formule sous modèle de Poisson par

$$\begin{aligned}&\mathbb{E}_{\mathcal{P}(z)}(L) = \sum_{k\geq 0}\sum_{i=0}^{k} z\binom{k}{i}p^i q^{k-i}\left(1-\exp\left(-p^i q^{k-i}z\right)\right) \textit{ et}\\ &\mathbb{E}_n(L) = \sum_{k\geq 0}\sum_{i=0}^{k} n\binom{k}{i}p^i q^{k-i}\left(1-\left(1-p^i q^{k-i}\right)^{n-1}\right) \textit{ sous modèle de Bernoulli.}\end{aligned} \tag{1.12}$$

La détermination de la moyenne de la taille se fait de la même manière que celle de la longueur de cheminement. Il faut néanmoins obtenir l'expression de la probabilité $\mathbb{P}(N_w = 0)$. Sous un modèle de Poisson, le résultat est facile : $\mathbb{P}_{\mathcal{P}(z)}(N_w = 0) = \exp(-zp_w)$. Quand le nombre de textes de l'ensemble est fixé, l'événement $\{N_w = 0\}$ signifie qu'aucune des n chaînes de l'ensemble de base ne commence par w, soit

$$\mathbb{P}_n(N_w = 0) = (1-p_w)^n.$$

Les expressions de la moyenne de la taille sous les différents modèles sont données par :

Lemme 2 *Pour une source symétrique et sous modèle de Poisson de paramètre z, la moyenne de la taille S d'un trie a pour expression*

$$\mathbb{E}_{\mathcal{P}(z)}(S) = \sum_{k\geq 0} 2^k \left(1 - \left(1 + 2^{-k}z\right)\exp\left(-2^{-k}z\right)\right) \qquad et$$
$$\mathbb{E}_n(S) = \sum_{k\geq 0} 2^k \left(1 - \left(1 + 2^{-k}(n-1)\right)\left(1 - 2^{-k}\right)^{n-1}\right)$$

dans un modèle de Bernoulli. Pour une source biaisée (p, q), la moyenne de la taille S d'un trie se formule sous modèle de Poisson par

$$\mathbb{E}_{\mathcal{P}(z)}(S) = \sum_{k\geq 0}\sum_{i=0}^{k} \binom{k}{i} \left(1 - \left(1 + p^i q^{k-i} z\right)\exp\left(-p^i q^{k-i} z\right)\right) \text{ et sous modèle de Bernoulli par}$$
$$\mathbb{E}_n(S) = \sum_{k\geq 0}\sum_{i=0}^{k} \binom{k}{i} \left(1 - \left(1 + p^i q^{k-i}(n-1)\right)\left(1 - p^i q^{k-i}\right)^{n-1}\right). \tag{1.13}$$

Remarque : Dans les deux lemmes précédents, les résultats pour le cas symétrique sont séparés du cas biaisé pour mettre en avant leur forme simple. On note cependant que les résultats dans le cas symétrique ne sont que la particularisation des résultats du cas biaisé à $p = q = 1/2$.

1.5 Transformée de Mellin et comportement asymptotique

Dans cette section, la définition et quelques propriétés de la transformée de Mellin sont brièvement rappelées. Sa capacité à obtenir le comportement asymptotique d'une fonction est mise en avant par des résultats théoriques (première section) et un exemple pratique (seconde section) sur le comportement asymptotique de la moyenne de la longueur de cheminement dans un trie sous modèle de Poisson et pour une source sans mémoire. La transformée de Mellin a été introduite dans le champ de l'analyse d'algorithme dès 1972 par de Bruijn, Knuth et Rice [dBKR72]. Pour une présentation plus complète des propriétés de la transformée de Mellin, l'ouvrage de Flajolet, Gourdon et Dumas [FGD95] est une référence. Le livre de Szpankowski [Szp01] y consacre un chapitre complet avec de nombreuses applications.

La puissance de la transformée de Mellin tient dans le passage qu'elle offre entre, d'une part le développement asymptotique d'une fonction f en zéro ou en l'infini (exposant des puissances, termes logarithmiques et coefficients associés) et d'autre part le comportement singulier (pôles, résidus, multiplicité) de la fonction complexe $f^\star$, transformée de Mellin de f. Plus précisément, chaque terme en $x^c \log^k x$ du développement asymptotique de f correspond à un pôle d'ordre $k+1$ de $f^\star$ en $s = -c$.

1.5.1 Propriétés de base de la transformée de Mellin

Pour une fonction $f :]0, \infty[\to \mathbb{C}$ localement intégrable sur $[0,\infty[$, la transformée de Mellin de f est définie par :

$$\mathcal{M}(f(x); s) := f^\star(s) := \int_0^\infty f(t) t^{s-1}\, dt.$$

La *bande fondamentale* de $f^\star$ est la plus grande bande ouverte du plan complexe dans laquelle $f^\star(s)$ converge. Elle est notée $\langle \alpha, \beta \rangle$ pour deux réels α et β. La bande $\langle \alpha, \beta \rangle$ est l'ensemble des compexes dont la partie réelle est comprise entre α et β. La transformée de Mellin est analytique dans sa bande fondamentale.

Les quelques propriétés suivantes sont faciles à montrer.

Lemme 3 *Soit f une fonction qui admet une transformée de Mellin $f^\star$ dans la bande fondamentale $\langle \alpha, \beta \rangle$. Soit $\mu > 0$,*

$$\mathcal{M}(f(\mu x); s) = \mu^{-s} f^\star(s) \text{ et } \mathcal{M}(\mu f(x); s) = \mu f^\star(s). \tag{1.14}$$

Par linéarité, la somme harmonique $F(z) := \sum_{k \in K} \lambda_k f(\mu_k z)$ se transforme en

$$F^\star(s) := \mathcal{M}\left(\sum_{k \in K} \lambda_k f(\mu_k x); s \right) = \left(\sum_{k \in K} \lambda_k \mu_k^{-s} \right) f^\star(s), \tag{1.15}$$

où K est un ensemble fini d'indices.

La série $\sum_{k \in K} \lambda_k \mu_k^{-s}$ est appelée série de Dirichlet *associée à la somme harmonique F.*

Dans le lemme précédent, pour que $F^\star$ soit bien définie, il faut que la série de Dirichlet **et** la transformée de Mellin $f^\star$ soient bien définies. Donc la bande fondamentale de F est l'intersection du domaine de convergence de la série de Dirichlet[14] (un demi-plan du plan complexe pour K infini et les μ_k croissants) et de la bande fondamentale de $f^\star$.

La *partie singulière* de la fonction f en s_0 est la somme sur les indices négatifs des termes du développement en série de Laurent[15] de f au voisinage de s_0 : si la fonction f est méromorphe en $s = s_0$ alors, f se développe en série de Laurent au voisinage de s_0 en :

$$f(s) = \sum_{k=-\infty}^{\infty} c_k (s - s_0)^k,$$

où tous les c_k sont nuls en-dessous d'un certain indice négatif. La partie singulière vaut alors :

$$\sum_{k=-\infty}^{-1} c_k (s - s_0)^k.$$

Définition 5 *Soit f méromorphe sur Ω et $\mathcal{S} \subseteq \Omega$ l'ensemble des pôles de f dans Ω. La partie singulière de f sur Ω est la somme des parties singulières de f en chacun des points de $\mathcal{S}$. Si E est la partie singulière de f sur Ω, on note :*

$$f \asymp E.$$

Nous écrivons maintenant le théorème qui permet de passer de renseignements complexes sur la transformée de Mellin $f^\star$ de f au comportement asymptotique de f en zéro et en l'infini. Il est couramment appelé théorème d'inversion de Mellin. Il existe une version « directe » qui permet d'obtenir le développement singulier de la transformée de Mellin à partir du comportement asymptotique de la fonction de base f. Cette version sera utilisée dans le chapitre 4.

[14] Johann Peter Gustav Lejeune Dirichlet, mathématicien (1805–1859)

[15] Matthieu Paul Hermann Laurent, mathématicien luxembourgeois (1841–1908)

Théorème 1 *Soit f une fonction continue sur $]0,+\infty[$ dont la transformée de Mellin $f^\star$ admet une bande fondamentale non vide $\langle\alpha,\beta\rangle$. On suppose*

- *que $f^\star$ admet un prolongement méromorphe sur $\langle\alpha,\gamma\rangle$ avec $\gamma>\beta$ et est analytique sur $\Re=\gamma$;*
- *qu'il existe un réel $\eta\in]\alpha,\beta[$, un entier $r>1$ et une suite réelle $(T_j)_{j\in\mathbb{N}}$ strictement croissante, divergeant vers l'infini tels que $f^\star(s)=O(|s|^{-r})$ sur la réunion des segments $\{s\in\mathbb{C}\,|\,\Re(s)\in[\eta,\gamma],\Im(s)=T_j\}$ quand j tend vers l'infini.*

Si la partie singulière de $f^\star$ sur $\langle\eta,\gamma\rangle$ vérifie

$$f^\star \asymp \sum_{(\zeta,k)\in A} d_{\zeta,k}\frac{1}{(s-\zeta)^k}$$

où A est l'ensemble des couples (ζ,k) où ζ est un pôle de $f^\star$ et k une puissance positive inférieure à l'ordre du pôle (ordre compris), alors au voisinage de l'infini, on a le développement asymptotique

$$f(x)=\sum_{(\zeta,k)\in A} d_{\zeta,k}\frac{(-1)^k}{(k-1)!}x^{-\zeta}(\log x)^{k-1}+O(x^{-\gamma}).$$

1.5.2 Comportement asymptotique de la longueur de cheminement

Les propriétés de la transformée de Mellin introduites au paragraphe précédent vont nous servir à retrouver le comportement asymptotique de la moyenne de la longueur de cheminement d'un trie construit à partir d'un nombre de chaînes suivant une loi de Poisson de paramètre z et sous un modèle de source sans mémoire (p,q).

La moyenne de la longueur de cheminement s'écrit (cf. lemme 1) en fonction du paramètre z de la loi de Poisson

$$f(z):=\mathbb{E}_{\mathcal{P}(z)}(L)=\sum_{k\geq 0}\sum_{i=0}^{k} z\binom{k}{i}p^iq^{k-i}\left(1-\exp\left(-zp^iq^{k-i}\right)\right). \tag{1.16}$$

La fonction de base de la somme est $g(z):=z(1-\exp(-z))$. Le résultat (1.15) du lemme 3 permet de trouver la transformée de Mellin de f à partir de celle de g et de la série de Dirichlet. Au voisinage de zéro, la fonction $g(z)$ se comporte en z^2. Pour que l'intégrale de $g(z)z^{s-1}$ converge au voisinage de zéro, il faut que $s-1+2=s+1>-1$, soit $s>-2$. Le comportement de g est en z au voisinage de l'infini. L'intégrale de $g(z)z^{s-1}$ converge donc au voisinage de l'infini pour $s<-1$. La bande fondamentale de la transformée de Mellin de g est donc $\langle-2,-1\rangle$. Une intégration par parties offre le résultat

$$g^\star(s)=-\Gamma(s+1),$$

où la fonction Γ est définie pour $\Re(s)>0$ par

$$\Gamma(s):=\int_0^\infty t^{s-1}e^{-t}dt. \tag{1.17}$$

L'équation (1.15) permet d'écrire la transformée de Mellin de f

$$f^\star(s)=\sum_{k\geq 0}\sum_{i=0}^{k}\binom{k}{i}(p^iq^{k-i})^{-s}g^\star(s). \tag{1.18}$$

La série de Dirichlet associée à la transformée $f^\star$ est

$$\sum_{k\geq 0}\sum_{i=0}^{k}\binom{k}{i}(p^i q^{k-i})^{-s} = \sum_{k\geq 0}(p^{-s}+q^{-s})^k.$$

La bande de convergence de la série est l'ensemble des s pour lesquels la somme de Dirichlet converge. Cette somme converge si $(p^{-s}+q^{-s}) < 1$. La bande de convergence est donc $\Re < -1$. La bande fondamentale de la transformée de Mellin de f est l'intersection de la bande fondamentale de $g^\star$ et de la bande de convergence de la série de Dirichlet d'où

$$f^\star(s) = -\frac{\Gamma(s+1)}{1-(p^{-s}+q^{-s})} \text{ pour } s \in \langle -2, -1\rangle.$$

Notre but est de connaître le comportement asymptotique de f au voisinage de l'infini. On cherche donc les pôles de $f^\star$ à droite de -1. Les pôles de la fonction Gamma sont les entiers négatifs (y compris zéro), tous pôles simples. Le seul pôle de la fonction $s \mapsto \Gamma(s+1)$ à droite de $\Re = -1$ est -1. Les zéros du dénominateur sont plus compliqués à déterminer. Il y a trivialement un zéro simple en -1 ce qui crée un pôle double de $f^\star$ en -1. Le développement singulier de la transformée de Mellin au voisinage de -1 s'écrit

$$-\frac{1}{(s+1)^2(p\log p + q\log q)} + \left(\frac{\gamma}{p\log p+q\log q} - \frac{p\log^2 p + q\log^2 q}{2(p\log p+q\log q)^2}\right)\frac{1}{s+1}.$$

Il y a deux autres types de racines de $1-(p^{-s}+q^{-s})$: celles avec $\Re = -1$ et celles avec $\Re > -1$. Les racines sur l'axe $\Re = -1$ s'écrivent $-1+i\alpha$. Cela nous permet d'écrire deux équations en regardant la partie réelle et la partie imaginaire de l'équation :

$$\begin{aligned} p\cos(\alpha\log p) + q\cos(\alpha\log q) = 1, \text{ et} \\ p\sin(\alpha\log p) + q\sin(\alpha\log q) = 0. \end{aligned} \tag{1.19}$$

Chacune des deux équations est élevée au carré avant de les sommer. Deux conditions apparaissent : $\alpha(\log p - \log q) = 2k\pi$ et $\alpha\log q = 2k'\pi$ où k et k' sont deux entiers relatifs non nuls. Ces deux conditions impliquent que

$$\log p/\log q \in \mathbb{Q},$$

la source est alors dite *périodique*.

Si la source est périodique alors les points

$$-1 + \frac{2ik\pi}{\log p - \log q}$$

en sont les pôles (simples pour $k \in \mathbb{Z}^\star$ et double pour $k=0$). Ce sont les uniques pôles de $f^\star$ sur l'axe $\Re = -1$. Ces pôles sont placés périodiquement le long de l'axe $\Re = -1$ ce qui justifie le qualificatif de *périodique* pour la source. La contribution de ces pôles au comportement asymptotique est très faible mais néanmoins linéaire alors que pour les sources *apériodiques*, il n'y a pas de pôle sur l'axe $\Re = -1$ et donc le seul terme linéaire est issu du pôle en -1.

Que la source soit périodique ou non, les autres pôles (éventuels) de la transformée sont placés au-delà de $\Re = -1$ et ne comptent que pour $o(n)$ dans le comportement asymptotique. Les pôles différents de -1 qui apparaissent sur l'axe apportent chacun une partie singulière :

$$\frac{\Gamma\left(\frac{2ik\pi}{\log p-\log q}\right)}{\left(s+1-\frac{2ik\pi}{\log p-\log q}\right)h}.$$

On applique maintenant le théorème 1, dit « inverse » qui nous donne le comportement asymptotique de la moyenne de la longueur de cheminement sous modèle de Poisson. Il existe une méthode dite de dépoissonisation qui permet de passer du résultat sous le modèle de Poisson au résultat sous le modèle de Bernoulli de paramètre n. Nous ne détaillons pas cette méthode ici et nous obtenons le comportement asymptotique de la longueur de cheminement et de la taille sous le modèle de Bernoulli :

Proposition 2 *Le comportement asymptotique de l'espérance de la longueur de cheminement pour une source sans mémoire symétrique vaut :*

$$\mathbb{E}_n(L) = \frac{n\log n}{\log 2} + n\left(\frac{1}{2} + \frac{\gamma}{\log 2}\right) - n\epsilon_1(n) + o(n)$$

où

$$\epsilon_1(n) := \frac{1}{\log 2}\sum_{k\in\mathbb{Z}^\star}\Gamma\left(-1 + \frac{2ik\pi}{\log 2}\right) n^{-\frac{2ik\pi}{\log 2}}$$

est une fonction oscillante autour de zéro de très faible amplitude (de l'ordre de 10^{-5}) et γ est la constante d'Euler.

Pour une source générale et sans mémoire (p, q), on obtient comme comportement asymptotique si la source est périodique et s'il existe un pôle différent de -1 sur l'axe $\Re = -1$:

$$\mathbb{E}_n(L) = \frac{n\log n}{h} + \left(\frac{\gamma}{h} + \frac{p\log^2 p + q\log^2 q}{2h^2}\right) n + n\epsilon_2(n) + o(n),$$

où $h := -p\log p - q\log q$ est l'entropie de la source et

$$\epsilon_2(n) := \frac{1}{h}\sum_{k\in\mathbb{Z}^\star}\Gamma\left(\frac{2ik\pi}{\log p - \log q}\right) n^{-\frac{2ik\pi}{\log p - \log q}}$$

est une fonction oscillant autour de zéro de très faible amplitude (de l'ordre de 10^{-5}). Et si la source est apériodique :

$$\mathbb{E}_n(L) = \frac{n\log n}{h} + \left(\frac{\gamma}{h} + \frac{p\log^2 p + q\log^2 q}{2h^2}\right) n + o(n).$$

La méthode qui nous a permis d'obtenir le comportement asymptotique de la longueur de cheminement est aussi applicable pour la taille :

Proposition 3 *Le comportement asymptotique de la moyenne de la taille sous un modèle de Bernoulli et pour une source sans mémoire symétrique est donné par*

$$\mathbb{E}_n(S) = \frac{n}{\log 2} - \frac{n}{\log^2 2}\sum_{k\in\mathbb{Z}^\star} 2ik\pi\Gamma\left(-1 + \frac{2ik\pi}{\log 2}\right) n^{-2ik\pi/\log 2} + o(n).$$

La somme sur les $k \in \mathbb{Z}^\star$ est une fonction oscillante de très faible amplitude, de l'ordre de 10^{-5}.

Pour une source générale et sans mémoire (p, q), on obtient comme comportement asymptotique

$$\mathbb{E}_n(S) = \frac{n}{h} - n\epsilon_3(n) + o(n),$$

où

$$\epsilon_3(n) := \frac{1}{h^2} \sum_{k \in \mathbb{Z}^\star} 2ik\pi\Gamma\left(\frac{2ik\pi}{\log p - \log q}\right) n^{-\frac{2ik\pi}{\log p - \log q}},$$

est un fonction oscillante autour de zéro de très faible amplitude (de l'ordre de 10^{-5}*). Si la source est périodique et a des pôles différents de* -1 *sur l'axe* $\Re = -1$*, et*

$$\mathbb{E}_n(s) = \frac{n}{h} + o(n)$$

sinon.

Chapitre 2

Taille et longueur de cheminement d'un arbre des suffixes

Nous obtenons le comportement asymptotique de la taille et de la longueur de cheminement dans un arbre des suffixes sous un modèle de source sans mémoire. Nous utilisons les méthodes de la combinatoire analytique. L'arbre des suffixes est construit sur les n premiers suffixes d'un texte engendré par une source sans mémoire. Le comportement asymptotique de la taille et de la longueur de cheminement dans un trie sous un modèle sans mémoire sont connus. D'abord nous donnons un expression asymptotique de la moyenne de la taille et de la longueur de cheminement dans un arbre des suffixes. Puis nous montrons que la différence entre ces expressions et les expressions respectives pour un trie est asymptotiquement faible ainsi les comportements asymptotiques de la taille et de la longueur de cheminement d'un arbre des suffixes sont proches de ceux d'un trie.

We derive the asymptotic behavior of the average size and path length for suffix trees under a memoryless source model. We use the methodology of analytic combinatorics. The suffix tree is built from the n first suffixes of a text T generated from a memoryless source. The average behavior of size and path length in trie is known for a long time. First, we obtain the asymptotic expression of the average of these two parameters for a suffix tree. Then we show that the difference between these expressions and their trie counterparts is asymptotically small. Hence the asymptotic behavior of average size and path length are close to those of the trie.

Sommaire

2.1 Introduction

De nombreuses applications de l'arbre des suffixes dans l'informatique ont été mises en avant dans le chapitre précédent. On cherche maintenant à quantifier les performances en moyenne de cette structure de données, en particulier de deux de ses paramètres les plus importants : la taille et la longueur de cheminement. Les textes sont émis par source sans mémoire.

Dans leur article [JS94], Jacquet et Szpankowski ont développé une méthode de « chaîne étalon » (*string ruler*) pour obtenir des résultats sur l'espérance de la taille dans un arbre des suffixes sous un modèle de source sans mémoire. Nous reprenons leur idée de base, à savoir que l'espérance de la taille et de la longueur de cheminement d'un trie et d'un arbre des suffixes se comportent asymptotiquement de manière similaire. Notre idée directrice est donc de regarder le comportement asymptotique de la différence entre les deux quantités (pour un trie et pour un arbre des suffixes). Le comportement asymptotique de l'espérance de la taille et de la longueur de cheminement dans un trie est connu depuis longtemps (cf. chapitre précédent). Cela nous permet d'obtenir les premiers termes du comportement asymptotique pour un arbre des suffixes.

Notre approche fait appel à des méthodes de combinatoire analytique. Elle est plus simple que celle de Jacquet et Szpankowski et permet des résultats asymptotiques plus précis sur le terme d'erreur. Son troisième avantage est sa robustesse à une généralisation du modèle de génération des symboles, même si dans ce chapitre le modèle est plus contraint que celui de [JS94].

Pour construire un arbre des suffixes, on se donne un texte infini T écrit sur un alphabet binaire $\mathcal{A} = \{\mathtt{0}, \mathtt{1}\}$. Les symboles sont produits par une source sans mémoire. La lettre `0` est émise par la source avec probabilité p et la lettre `1` avec probabilité q. La probabilité d'émission p_w d'un texte qui commence par le motif w est le produit des probabilités des lettres qui composent w. On adopte la convention $p \geq q$.

Dans notre modèle, le nombre de chaînes à partir desquelles est formé l'arbre des suffixes est fixé et vaut n (contrairement au modèle de Poisson que nous avons utilisé dans le chapitre précédent). On rappelle que le paramètre $\widehat{N}_w(T, n)$ est le nombre de chaînes parmi les n premiers suffixes du texte T qui commencent par w. On dit que le motif w apparaît à la position j dans le texte T si $T_j \dots T_{j+|w|-1} = w$. Le paramètre $\widehat{N}_w(T, n)$ est aussi le nombre d'apparitions du motif w dans les n premières positions du texte T.

Nous obtenons le comportement asymptotique de l'espérance de la taille moyenne (resp. de la longueur de cheminement) d'un arbre des suffixes construit sur les n premiers suffixes d'un texte engendré par un modèle de source sans mémoire. Le théorème suivant résume nos résultats. Il a été publié dans [Fay04].

Théorème *Le comportement asymptotique de l'espérance de la taille et de la longueur de cheminement d'un arbre des suffixes construit sur les n premiers suffixes d'un mot produit par une source sans mémoire (p, q) avec $p \geq q$ et $0.5 \leq p \leq 0.54$ sont donnés par :*

$$\mathbb{E}_n(S) = \frac{n}{h}(1 + \epsilon_1(n)) + O(n^{0.85}),$$
$$\mathbb{E}_n(L) = \frac{n \log n}{h} + \left(\frac{\gamma}{h} + \frac{p \log^2 p + q \log^2 q}{2h^2} + \epsilon_2(n)\right) n + O(n^{0.85}),$$

où $h = -p \log p - q \log q$ est l'entropie de la source, γ est la constante d'Euler et les ϵ_i sont des fonctions fluctuant autour de zéro de très faible amplitude (typiquement inférieure à 10^{-5}).

L'expression de l'espérance de la taille et de la longueur de cheminement dans un arbre des suffixes fait intervenir les probabilités $\mathbb{P}_n(\widehat{N}_w = 0)$ et $\mathbb{P}_n(\widehat{N}_w = 1)$ qu'un texte de taille n ne contienne aucune occurrence de w et qu'un texte de taille n contienne une unique occurrence de w. Dans la section 2.2 on obtient par deux méthodes les séries génératrices des probabilités des textes avec un nombre donné d'occurrences (ici 1 et 0) d'un certain motif w fixé.

La section 2.3 vise à trouver les comportements asymptotiques des probabilités $\mathbb{P}_n(\widehat{N}_w = 0)$ et $\mathbb{P}_n(\widehat{N}_w = 1)$. On se sert pour cela des techniques de combinatoire analytique (*analytic combinatorics*). Nous cherchons les pôles dominants des séries génératrices qui comptent les

textes contraints par le nombre d'occurrences de w. Le théorème de Rouché appliqué dans la section 2.3.2 affirme l'unicité du pôle dominant (lemme 7). Le théorème de Cauchy utilisé dans la section 2.3.3 offre le comportement asymptotique du coefficient d'ordre n des séries génératrices de probabilités, c'est-à-dire les deux probabilités recherchées $\mathbb{P}_n(\widehat{N}_w = 0)$ et $\mathbb{P}_n(\widehat{N}_w = 1)$.

Dans la section 2.4, l'asymptotique de la différence entre l'espérance de la longueur de cheminement dans un arbre des suffixes et l'espérance de la longueur de cheminement dans un trie construit sur un ensemble poissonisé (c'est-à-dire que le cardinal suit une loi de Poisson) est étudiée. Cette différence est asymptotiquement faible. Pour prouver cela, l'étude est découpée en plusieurs sections selon la taille des motifs : les motifs courts (section 2.4.1), les motifs longs (section 2.4.2) et deux types de motifs intermédiaires (section 2.4.3) les périodiques et les apériodiques. La section 2.5 montre que la moyenne de la taille d'un trie et d'un arbre des suffixes se comporte asymptotiquement de manière identique. Elle est bâtie sur le même principe que la section 2.4.

Dans le chapitre précédent, le comportement asymptotique de la moyenne de la taille et de la longueur de cheminement dans un trie ont été rappelés. Il en résulte une caractérisation précise du comportement asymptotique de la moyenne de la taille et de la longueur de cheminement dans un arbre des suffixes construit sur les n premiers suffixes d'un texte engendré par une source sans mémoire.

La notion clef de l'étude est celle d'auto-corrélation, c'est-à-dire la faculté structurelle qu'a un mot w de pouvoir réapparaître rapidement dans un texte. Le codage de cette propriété s'effectue par un polynôme d'auto-corrélation $c_w(z)$ qui va jouer un rôle majeur dans l'expression des probabilités et dans les majorations des deux dernières sections.

Une fois établi, le résultat du théorème sur la longueur de cheminement permet de comprendre l'efficacité de l'algorithme de compression LZ'77 expliqué dans la section 1.3 et basé sur les arbres des suffixes. Les n premiers suffixes du texte ont été insérés dans l'arbre des suffixes et la longueur de cheminement vaut en moyenne $(1/h)n \log n$. Chacun des n suffixes crée une unique feuille et la profondeur moyenne de chacune de ces feuilles est $(1/h) \log n$.

Le principe de l'algorithme LZ'77 est de construire l'arbre des suffixes du texte, de trouver le mot le plus long w dans l'arbre des suffixes (mot déjà vu) qui correspond au mot lu à partir de la position courante. Ensuite l'occurrence courante de w dans le texte initial est remplacée par le numéro de son nœud associé dans l'arbre des suffixes pour obtenir le texte compressé. Il y a en moyenne n/h nœuds dans l'arbre des suffixes donc les indices de ces nœuds sont au plus de longueur $\log n$. On remarque que les nœuds avec deux fils non vides ne peuvent pas correspondre à des plus longs motifs rencontrés dans le texte. Seuls les feuilles et les nœuds unaires peuvent correspondre à un plus long motif déjà rencontré. Pour les feuilles, leur profondeur est en moyenne $(1/h) \log n$. Pour résumer, le texte est divisé en motif u de longueur moyenne $(1/h) \log n$ et chacun de ces motifs est remplacé dans le texte compressé par un numéro de longueur $\log n$. Le texte compressé fait une taille de hn et on retrouve le taux de compression h.

2.2 Détermination des séries génératrices comptant le nombre d'occurrences d'un motif

Les séries génératrices qui pesent les textes avec aucune (resp. une unique) occurrence du motif w dans les n premières positions du texte sont notées $\mathfrak{N}_w(z)$ (resp. $\mathfrak{O}_w(z)$). Nous obtenons leur expression dans cette section.

La première intuition est d'écrire que les textes avec une seule occurrence de w se décomposent

comme $\mathcal{A}^*.w.\mathcal{A}^*$. Las ! les choses ne sont pas aussi simples ! D'abord rien n'interdit que l'un des textes de l'ensemble $\mathcal{A}^*$ ne contienne pas d'apparition du motif w. D'autre part le motif w peut apparaître entre les deux ensembles, c'est-à-dire « à cheval » sur l'occurrence de w visible dans la décomposition et sur un mot de $\mathcal{A}^*$.

Regardons ce qui se passe avec le motif $w = \texttt{000000}$. Ce motif est très particulier mais il fait ressortir de manière frappante le concept d'**auto-corrélation**. Nous cherchons à compter le nombre d'apparitions du motif w dans le texte T. Une fois que l'on a trouvé w dans le texte (disons à la position j) il suffit que la lettre à la position $j + |w| = j + 6$ soit `0` pour retrouver une nouvelle occurrence de w. L'auto-corrélation rend compte de la manière dont l'apparition du motif w dans un texte peut renseigner sur une apparition ultérieure du même motif.

2.2.1 Auto-corrélation et méthode de Guibas et Odlyzko

La série génératrice $\mathfrak{N}_w(z)$ des textes sans aucune occurrence du motif w de taille k est déterminée dans cette section par une méthode originellement développée par Guibas et Odlyzko [GO81b]. Elle est définie par

$$\mathfrak{N}_w(z) := \sum_{n\geq 0} \mathbb{P}_n(\widehat{N}_w = 0) z^n. \tag{2.1}$$

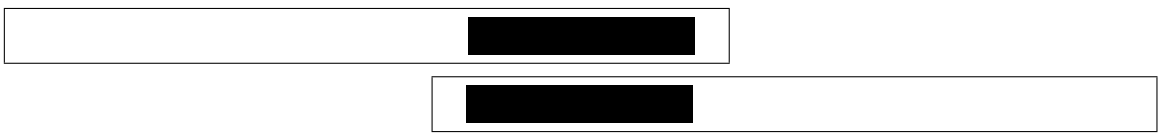
Prenons un texte T sans aucune occurrence de w et rajoutons-lui une lettre à la fin. Soit cette lettre va créer une occurrence finale de w (et ce sera la seule occurrence de w dans le nouveau texte étendu), soit cette nouvelle lettre ne va créer aucune occurrence de w. Cette décomposition est valable pour n'importe quel texte T sans aucune occurrence de w et elle est non-ambiguë. Si on introduit les ensembles des textes avec aucune occurrence de w (noté $\mathcal{N}$) et avec une unique occurrence finale de w (noté $\mathcal{T}$), on récrit cette décomposition sous la forme

$$\mathcal{N}.\mathcal{A} + \epsilon = \mathcal{T} + \mathcal{N}. \tag{2.2}$$

On doit rajouter le mot vide ϵ puisqu'il appartient à l'ensemble $\mathcal{N}$, mais ne peut appartenir à $\mathcal{N}.\mathcal{A}$.

Reprenons un texte T sans aucune occurrence de w, et ajoutons-lui le motif w à la fin. Réutilisons le motif « pathologique » $w = \texttt{0000}$ pour bien faire comprendre le phénomène : si le texte T se termine par `0` et qu'on lui rajoute w à la fin, alors on va créer une occurrence de w qui n'est pas finale. Pour obtenir une équation aussi élégante que la précédente, on doit prendre en compte la possibilité qu'une occurrence de w se crée avant l'occurrence finale. Que peut-il se passer ? Une occurrence de w peut arriver dans le texte $T.w$ avant l'occurrence finale si T finit par un préfixe de w (disons celui de taille j) et le mot w (de taille k) que l'on ajoute va permettre de finir cette occurrence inachevée de w. Il faut pour cela que le préfixe de taille $k - j$ de w, soit aussi le suffixe de taille $k - j$ de w. Alors $T.w$ se décompose alors en un texte U avec une occurrence unique et finale de w et le suffixe v de taille j de w.

La condition pour que w apparaisse dans $T.w$ avant l'occurrence finale correspond à la possibilité qu'un suffixe et un préfixe de même taille $(k - j)$ soient identiques, c'est-à-dire un chevauchement de motif w avec lui-même de la forme :



où la partie en noir est le préfixe/suffixe de taille $k - j$.

L'ensemble $\mathcal{C}_w$ des suffixes v qui complètent le texte U en donnant lieu à une nouvelle occurrence de w est défini par :

$$\mathcal{C}_w = \{v : \exists u \neq \epsilon, v' : w = v'.u = u.v\}.$$

La décomposition que nous venons de proposer est valable pour tous les textes sans aucune occurrence de w, et on obtient l'équation

$$\mathcal{N}.w = \mathcal{T}.\mathcal{C}_w. \tag{2.3}$$

La série génératrice de l'ensemble $\mathcal{C}$ est notée $C(z)$ et $c(z)$ est la série génératrice probabilisée de l'ensemble $\mathcal{C}$. La notation $C(z)$ ne servira presque jamais, il n'y donc aucun risque de confusion entre les différents « C. » L'ensemble $\mathcal{C}$ contient au plus $|w| - 1$ éléments et est donc fini, ainsi les séries génératrices sont sur un nombre fini de termes, c'est-à-dire des polynômes.

Définition 6 *Le polynôme d'auto-corrélation probabilisé $c_w(z)$ du motif w est défini par :*

$$c_w(z) := \sum_{j=0}^{k-1} c_{j,w} z^j \mathbb{P}(w_{k-j+1} \dots w_k), \tag{2.4}$$

où $c_{j,w}$ vaut 1 si le suffixe et le préfixe de taille $k-j$ de w sont identiques et zéro sinon.

Remarque : Autant que possible les indices w seront omis que ce soit sur les coefficients $c_{j.w}$ ou sur le polynôme d'auto-corrélation c_w.

Dans le cas d'une source symétrique, le polynôme d'auto-corrélation probabilisé se particularise en :

$$c(z) = \sum_{j=0}^{k-1} c_j z^j \frac{1}{2^j} = \sum_{j=0}^{k-1} c_j \left(\frac{z}{2}\right)^j = C\left(\frac{z}{2}\right).$$

Cette expression n'est valable que pour une source symétrique.

Il existe un dictionnaire nous permettant de passer des ensembles aux séries génératrices qui comptent les éléments de ces ensembles selon un (ou plusieurs) paramètres. On passe du système sur les langages à un système sur les séries génératrices des probabilités :

$$\begin{cases} \mathcal{N}.\mathcal{A} + \epsilon = \mathcal{T} + \mathcal{N} \\ \mathcal{N}.w = \mathcal{T}.\mathcal{C} \end{cases} \longleftrightarrow \begin{cases} \mathfrak{N}_w(z)z + 1 = \mathfrak{T}_w(z) + \mathfrak{N}_w(z) \\ \mathfrak{N}_w(z)p_w z^k = \mathfrak{T}_w(z)c_w(z) \end{cases}, \tag{2.5}$$

où $\mathfrak{T}_w(z)$ est la série génératrice probabilisée associée à l'ensemble des textes $\mathcal{T}$.

Le système de droite (équations sur les séries génératrices) se résout de manière classique et amène l'équation

$$\mathfrak{N}_w(z) = \sum_{n\geq 0} \mathbb{P}_n(\widehat{N}_w = 0) z^n = \frac{c(z)}{c(z)(1-z) + p^i q^{k-i} z^k}, \tag{2.6}$$

où i est le nombre d'occurrences de la lettre `0` dans le motif w.

Il existe un système du même type (quoique légèrement plus compliqué) qui fait intervenir l'ensemble des textes $\mathcal{O}$ qui ne contiennent qu'une seule occurrence du motif w. La série génératrice probabilisée $\mathfrak{O}(z)$ qui compte les textes avec une unique apparition de w est déterminée explicitement à partir de ce système. Elle est définie par

$$\mathfrak{O}(z) := \sum_{n\geq 0} \mathbb{P}_n(\widehat{N}_w = 1) z^n. \tag{2.7}$$

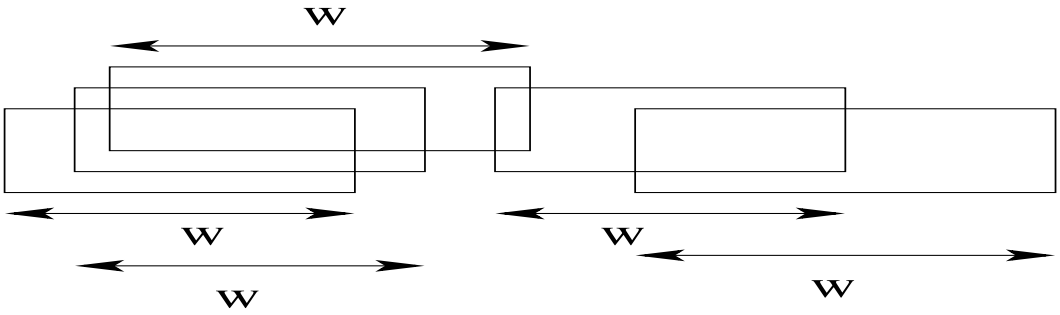


FIG. 2.1 – Un amas de motifs w.

Il existe néanmoins une autre méthode plus générale pour exprimer ces séries génératrices avec une contrainte sur le nombre d'apparition du motif w : trouver la série génératrice bivariée $W_w(z,u)$ qui compte les textes selon le nombre d'occurrences du motif w (par la variable u) et selon la taille du texte (par la variable z). Cette méthode s'appelle l'inclusion-exclusion.

2.2.2 Méthode d'inclusion-exclusion

Le principe d'inclusion-exclusion a été introduit pour le comptage de textes contraints par certains conditions par Goulden et Jackson [GJ83] et reformulé dans [FS06]. Il permet de déterminer la série génératrice bivariée $W_w(z,u)$ du nombre d'occurrences d'un motif w fixé en marquant par u son nombre d'occurrences et par z la taille du texte. Le procédé est expliqué brièvement ici avant d'obtenir l'expression de la série génératrice bivariée $W_w(z,u)$.

Au lieu de compter individuellement chaque occurrence du motif w dans le texte T, on compte les *amas* d'occurrences de w. Un amas (dit parfois *grappe* ou *clique* dans la littérature francophone et *cluster* ou *clump* en anglais) est un texte composé uniquement par des motifs w. S'il y a plus de deux motifs w dans cet amas, chaque motif w en chevauche au moins un autre (en une position quelconque). On a en quelque sorte un empilement de motifs w (cf. figure 2.1). L'ensemble des amas de w est noté $\mathcal{K}_w$ ou plus simplement $\mathcal{K}$.

On change l'alphabet de $\mathcal{A}$ en $\{\mathcal{A},\mathcal{K}\}$ pour décomposer le texte T. Le texte est alors la concaténation de lettres de l'alphabet $\mathcal{A}$ et d'amas de $\mathcal{K}$. Cette décomposition n'est pas unique. On définit l'ensemble $\mathcal{F} := (\mathcal{A}+\mathcal{K})^\star$. La série génératrice $K(z,v)$ de l'ensemble $\mathcal{K}$, où z marque la taille de l'amas et v le nombre d'occurrences de w dans l'amas, se détermine en décomposant l'ensemble $\mathcal{K}$

$$\mathcal{K}_w = w + w.(\mathcal{C}_w - \epsilon) + w.(\mathcal{C}_w - \epsilon)^2 + \cdots .$$

Le dictionnaire entre langages et séries génératrices permet d'obtenir

$$K(z,v) = \frac{vz^{|w|}p_w}{1 - v(c(z)-1)} \quad \text{et} \quad F(z,v) = \frac{1}{1 - z - K(z,v)},$$

où z marque la taille du texte (en lettres de $\mathcal{A}$) et v le nombre d'occurrences de w dans les amas. On trouve la série génératrice $W_w(z,u)$ en changeant la variable v en $u-1$. Cela permet de compter toutes les occurrences de w dans le texte par la variable u. Expliquons brièvement ce changement de variables dans le sens « inverse » (changement de u en $1+v$). Dans un texte où les occurrences de w sont marquées par u, on change de variable de marquage : certaines occurrences de w dans le textes sont marquées par la variable v et d'autres par la variable 1 (ce qui revient *de facto* à masquer ces occurrences). On obtient alors un texte avec des amas d'occurrences de w (amas éventuellement formé d'une seule occurrence) et des lettres « libres »

de $\mathcal{A}$ (qui peuvent éventuellement contenir des occurrences masquées de w), c'est-à-dire un texte de $\mathcal{F}$.

L'application de la méthode d'inclusion-exclusion amène à l'expression de la série génératrice de probabilités bivariée :

$$\boxed{W_w(z,u) = \sum_{l,m\geq 0} \mathbb{P}_m(\widehat{N}_w = l)u^l z^m = \frac{(u-1)c(z)-u}{(1-z)((u-1)c(z)-u)+(u-1)z^k p_w}.} \tag{2.8}$$

L'extraction du terme en u de $W(z,u)$ apporte la série génératrice $\mathfrak{O}(z)$. En extrayant le coefficient z^0, on retrouve bien l'expression de $\mathfrak{N}(z)$. Soit

Lemme 4 *Les séries génératrices $\mathfrak{N}_w(z)$ (resp. $\mathfrak{O}_w(z)$) qui comptent les textes avec aucune (resp. une unique) occurrence du motif w dans les n premières positions du texte s'écrivent*

$$\mathfrak{N}_w(z) = \sum_{n\geq 0} \mathbb{P}_n(\widehat{N}_w = 0)z^n = \frac{c(z)}{c(z)(1-z)+p^i q^{k-i} z^k} = \frac{c(z)}{c(z)(1-z)+p_w z^k} \text{ et}$$

$$\mathfrak{O}_w(z) = \sum_{n\geq 0} \mathbb{P}_n(\widehat{N}_w = 1)z^n = \frac{z^k p^i q^{k-i}}{(c(z)(1-z)+p^i q^{k-i} z^k)^2} = \frac{z^k p_w}{(c(z)(1-z)+p_w z^k)^2}, \tag{2.9}$$

où i est le nombre d'occurrences de la lettre `0` *dans le motif w.*

2.2.3 Propriété des valeurs du polynôme d'auto-corrélation en 1

Dans cette courte section, nous montrons une propriété des valeurs en 1 du polynôme d'auto-corrélation probabilisé. Celle-ci sera très utile pour obtenir le lemme 11.

Lemme 5 *La somme sur tous les motifs d'une taille donnée des valeurs en 1 du polynôme d'auto-corrélation probabilisé dans le cas d'une source sans mémoire ne dépend pas des valeurs de la probabilité p et vaut*

$$\sum_{w\in\mathcal{A}^k} c_w(1) = 2^k + k - 1. \tag{2.10}$$

Preuve : Sur un alphabet binaire, il y a 2^j motifs de taille j pour n'importe quel $1 \leq j < k$. Pour chaque motif v de taille j, on construit un unique mot w de taille k pour lequel $c_{j,w} = 1$ et v est son suffixe de taille j (et on construit ce mot w en répétant le motif v autant de fois que possible). Il y a donc *au plus* 2^j motifs de taille k avec un suffixe de taille j donné et qui satisfont $c_j = 1$. De plus, on ne peut en avoir moins de 2^j, sinon il y aurait deux motifs de taille j différents qui créeraient le même mot w. Il y a donc *exactement* 2^j motifs de taille k qui satisfont $c_j = 1$ et ce pour chaque j entre 1 et $k-1$. D'où

$$\begin{aligned}\sum_{w\in\mathcal{A}^k} c_w(1) &= \sum_{w\in\mathcal{A}^k}\sum_{j=0}^{k-1} c_{j,w}\mathbb{P}(w_{k-j+1}\cdots w_k) = \sum_{j=0}^{k-1}\sum_{w\in\mathcal{A}^k:\, c_{j,w}=1} \mathbb{P}(w_{k-j+1}\cdots w_k)\\ &= 2^k + \sum_{j=1}^{k-1}\sum_{v\in\mathcal{A}^j}\mathbb{P}(v_1\cdots v_j) = 2^k + \sum_{j=1}^{k-1} 1 = 2^k + k - 1,\end{aligned}$$

puisque nous venons de montrer qu'à chaque w de taille k avec $c_j = 1$, il existe un unique v de taille j, qui est en fait le suffixe de taille j de w ◄

2.3 Asymptotique des coefficients

Un théorème des mathématiques affirme que le comportement asymptotique du coefficient d'ordre n d'une série génératrice rationnelle $G(z)$ se comporte en $|\rho|^{-n}$ où ρ est le pôle dominant de la série génératrice (si ce pôle est unique). On cherche donc à localiser précisément les pôles de plus petit module de $\mathfrak{N}(z)$ et $\mathfrak{O}(z)$, car ce sont ces pôles qui déterminent le comportement asymptotique des coefficients des deux séries génératrices. On procède en plusieurs étapes : d'abord le théorème de Pringsheim garantit qu'il existe un pôle dominant sur l'axe réel positif. On en trouve ensuite une valeur approchée, c'est-à-dire la valeur approchée de la racine réelle de plus petit module du dénominateur

$$\mathfrak{D}_w(z) = z^k p^i q^{k-i} + (1-z)c(z) \tag{2.11}$$

des deux séries génératrices. Ensuite le théorème de Rouché permet de montrer que ce pôle est l'unique pôle dominant des deux séries génératrices. On conclut en appliquant le théorème de Cauchy aux séries génératrices sur un contour circulaire qui nous donne les comportements asymptotiques de $\mathbb{P}_n(\widehat{N}_w = 0)$ et $\mathbb{P}_n(\widehat{N}_w = 1)$ quand n tend vers l'infini.

Le théorème de Pringsheim affirme qu'une des singularités de plus petit module d'une série génératrice à coefficients positifs et à rayon de convergence fini est réelle et positive. Les séries génératrices $\mathfrak{N}(z)$ et $\mathfrak{O}(z)$ sont à coefficients positifs, donc elles ont une singularité réelle dominante positive que l'on note ρ. C'est la racine réelle de plus petit module de $\mathfrak{D}(z)$. Nous cherchons maintenant les zéros dominants de $\mathfrak{D}(z)$.

2.3.1 Résolution approchée

On détermine dans cette section une valeur approchée utilisable pour ρ, le zéro réel positif de plus petit module de $\mathfrak{D}(z) = z^k p^i q^{k-i} + (1-z)c(z)$.

Remarquons d'abord que $\mathfrak{D}(1) = p^i q^{k-i}$. Cette quantité est proche de zéro (si on a un motif composé de répétitions de la même lettre, cette lettre ayant en plus une probabilité proche de 1, la décroissance exponentielle est faible mais elle reste exponentielle). Pour compenser le terme $p^i q^{k-i}$ crée par l'approximation à l'ordre 0, on introduit η_1, terme de correction à l'ordre 1 de la racine réelle dominante. Il vérifie $p^i q^{k-i} - \eta_1 c(1+\eta_1) = 0$, soit

$$\eta_1 = \frac{p^i q^{k-i}}{c(1+\eta_1)}.$$

Sur l'axe réel positif, le polynôme d'auto-corrélation $c(z)$ est plus grand que 1. Il est borné par une constante pour z suffisamment près de 1. Donc $\eta_1 = \Theta(p_w)$.

La racine ρ s'écrit sous la forme $1+\eta$ et on va en effectuer un développement selon les puissances de p_w. L'approximation η_1 de η, est de l'ordre de p_w, ainsi pour des motifs de grande taille

$$\begin{aligned} 0 = \mathfrak{D}(\rho) = \mathfrak{D}(1+\eta) &= p_w(1+\eta)^k - \eta c(1+\eta) \\ &= p_w(1+k\eta+O(\eta^2)) - \eta(c(1)+\eta c'(1)+O(\eta^2)) = (p_w - \eta c(1)) + O(p_w^2). \end{aligned}$$

Lemme 6 *Le pôle réel dominant de la fonction $\mathfrak{D}_w(z)$ vérifie*

$$\rho = 1 + \frac{p_w}{c(1)} + O(p_w^2),$$

quand la taille de w tend vers l'infini.

2.3.2 Unicité du pôle dominant

Le réel ρ a été défini comme la solution dominante réelle de l'équation $\mathfrak{D}(z) = 0$. On montre en utilisant le théorème de Rouché que ρ est l'unique zéro dominant de $\mathfrak{D}(z)$. Il est donc l'unique pôle dominant des séries génératrices $\mathfrak{N}(z)$ et $\mathfrak{O}(z)$. Nous allons montrer le lemme suivant

Lemme 7 *Pour une source sans mémoire (p, q) et tout réel $\phi > 1$ avec $p\phi < 1$, il existe un entier K, tel que dans le disque centré de rayon ϕ et pour les motifs de taille supérieure à K, la fonction $\mathfrak{D}_w(z) = p_w z^k + (1 - z)c(z)$ ne s'annule qu'une seule fois.*

Nous avons besoin pour démontrer le lemme précédent du théorème de Rouché que nous rappelons :

Théorème 2 (Rouché) *Soit γ une courbe simple, fermée. Supposons que dans le domaine fermé délimité par la courbe γ, les fonctions f et g soient analytiques et qu'elles satisfassent : $|g(z)| < |f(z)|$ pour $z \in \gamma$. Alors $f(z)$ et $f(z) + g(z)$ ont le même nombre de zéros à l'intérieur du domaine délimité par γ.*

Le théorème de Rouché est appliqué à un cercle $\mathfrak{C}$ de centre l'origine et de rayon R. La fonction la plus susceptible d'être dominée est $g_w(z) := z^k p^i q^{k-i}$ et on prend $f_w(z) := c_w(z)(1 - z)$. Les deux fonctions f et g (on se débarrasse des indices w) sont des polynômes, donc analytiques sur le cercle $\mathfrak{C}$ et aussi dans le disque délimité par $\mathfrak{C}$ comme le requiert le théorème. Pour pouvoir appliquer le théorème de Rouché, nous cherchons un rayon R du cercle $\mathfrak{C}$ tel que $|g(z)| < |f(z)|$ si $|z| = R$. Pour cela le lemme préliminaire suivant est important.

Lemme 8 *Pour tout $\phi > 1$ tel que $p\phi < 1$, il existe K et $\alpha > 0$ tels que, pour chaque motif w de taille plus grande que K et pour z dans le disque de rayon ϕ, on ait*

$$|c(z)| \geq \alpha.$$

Remarque : Ce lemme est une version adaptée au cas des sources sans mémoire du lemme 2 de Fayolle et Ward [FW05]. Dans l'article original les sources sont de type markovien, on y reviendra dans le lemme 42 du chapitre 5 (page 137).

Preuve : Soit w un motif de taille k, $\phi > 1$ qui vérifie $p\phi < 1$ et i la plus petite puissance strictement positive dans le polynôme d'auto-corrélation. La preuve se divise en deux parties selon que l'indice i est plus petit ou plus grand que $k/2$. Le polynôme d'auto-corrélation vérifie toujours $c_0 = 1$ puisque le motif se chevauche toujours avec lui-même ($w = w.\epsilon = \epsilon.w$), et on écrit donc

$$c_w(z) = 1 + \sum_{j=i}^{k-1} c_j \mathbb{P}(w_{k-j+1}^k) z^j. \tag{2.12}$$

Si dans le polynôme d'auto-corrélation $c_j = 1$, le suffixe u de taille j de w va déterminer complètement le motif w en se répétant par translation. On regarde la figure 2.2 pour bien saisir ce phénomène : on a égalité entre les lettres de la configuration A (on parle de configuration car il s'agit du même motif w) et celles de la configuration B qui sont juste en-dessous. Cela se traduit de manière plus analytique par $w_i = w_{i-j}$ (pour une certaine plage d'indices i). La même chose est vraie entre la configuration B et la configuration C. Mais par transitivité on a aussi une relation d'égalité entre deux lettres de w à distance $2j$, c'est-à-dire une relation entre les configurations A et C.

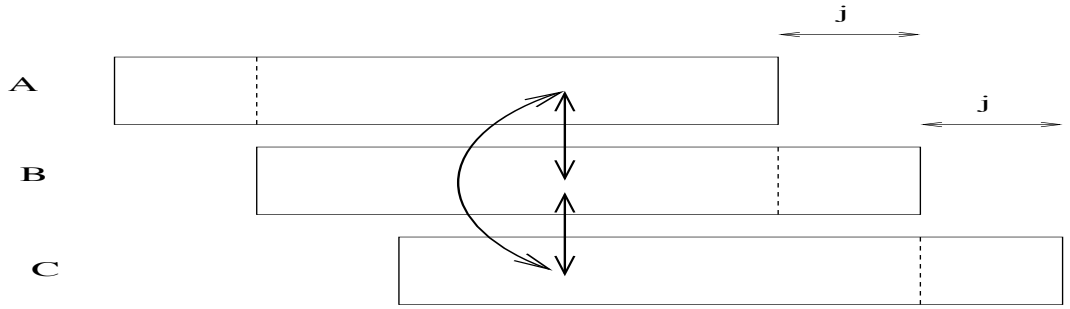


FIG. 2.2 – Le motif w, le même décalé de j puis encore de j.

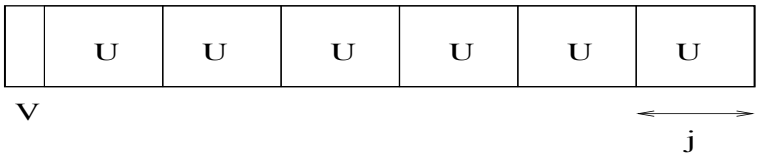


FIG. 2.3 – Si $c_i = 1$ alors w s'écrit en fonction de ses i dernières lettres : $w = v.u \cdots u$, où v est le suffixe de u qui fait le reste.

Le motif w va donc s'écrire comme $q := \lfloor k/i \rfloor$ fois le motif u et, si w ne peut s'écrire comme un nombre entier de répétitions de u, alors le suffixe v de u de taille $r := k - \lfloor k/i \rfloor i$ va se retrouver en préfixe de w pour compléter l'écriture de w (cf. figure 2.3). On introduit aussi le mot v' tel que $v'v = u$.

Si $i > \lfloor k/2 \rfloor$, il n'y a aucune répétition du motif u et le mot w se décompose en $v.u$. On a les minorations suivantes

$$|c_w(z)| \geq 1 - \left| \sum_{j=i}^{k-1} c_j \mathbb{P}(w_{k-j+1}^k) z^j \right| \geq 1 - \frac{(p\phi)^i}{1-p\phi}, \tag{2.13}$$

pour $|z| \leq \phi$. Mais puisque $i > \lfloor k/2 \rfloor$ et $p\phi < 1$, on obtient

$$|c_w(z)| \geq 1 - \frac{(p\phi)^{k/2}}{1-p\phi}.$$

Nous remarquons que pour une longueur K_1 suffisamment grande, tous les motifs w de taille plus grande que K_1 vont vérifier $|c_w(z)| \geq \alpha$ pour un α positif.

Si $i \leq \lfloor k/2 \rfloor$, le suffixe de taille i va se répéter. Le polynôme d'auto-corrélation s'écrit alors :

$$c_w(z) = 1 + \mathbb{P}(u) z^i + \mathbb{P}(u.u) z^{2i} + \cdots + \mathbb{P}(u^{q-1}) z^{i(q-1)} c_{uv}(z).$$

Il ne peut y avoir d'autres monômes que les multiples de i dans l'écriture du polynôme d'auto-corrélation. Si on avait un indice l entre i et $2i$ (par exemple) alors le motif w se superposerait avec le motif décalé de $l - i$, et cela contredirait l'hypothèse de minimalité de i. Néanmoins les relations sont beaucoup plus complexes pour les corrélations à la fin du mot w. Pour éviter de rentrer dans ces questions, on écrit juste $c_{uv}(z)$ et cela nous suffira.

On est dans un modèle de source sans mémoire donc $\mathbb{P}(u^{q-1}) = \mathbb{P}(u)^{q-1}$, et ainsi

$$c_w(z) = 1 + p_u z^i + (p_u z^i)^2 + \cdots + (p_u z^i)^{q-1} c_{uv}(z) = \frac{1 - (p_u z^i)^{q-1}}{1 - p_u z^i} + (p_u z^i)^{q-1} c_{uv}(z).$$

Nous minorons ensuite $|c_w(z)|$ pour $|z| \leq \phi$:

$$|c_w(z)| \geq \left|\frac{1-(p_u z^i)^{q-1}}{1-p_u z^i}\right| - \left|(p_u z^i)^{q-1} c_{uv}(z)\right| \geq \frac{1-(p\phi)^{i(q-1)}}{1+(p\phi)^i} - (p\phi)^{i(q-1)}|c_{uv}(z)|$$
$$\geq \frac{1-(p\phi)^{i(q-1)}}{1+(p\phi)^i} - \frac{(p\phi)^{i(q-1)}}{1-p\phi}.$$

On se sert de la majoration de p_u par p^i où i est la taille de u et du fait que $|c_{uv}(z)| \leq (1-p\phi)^{-1}$ pour $|z| \leq \phi$.

À partir de ce résultat, l'hypothèse $p\phi < 1$ nous garantit que $(p\phi)^k$ tend vers zéro quand k tend vers l'infini. Le produit $i(q-1)$ est proche de k (au pire il vaut $k/3$ si $w = uuv$) et donc pour les motifs d'une taille suffisamment grande (disons d'une taille supérieure à un certain K_2), le terme $(1+(p\phi)^i)^{-1}$ domine. Cela signifie que nous avons minoré le polynôme d'auto-corrélation sur le cercle de rayon ϕ par une quantité strictement positive. On conclut en posant $K = \max\{K_1, K_2\}$. ◄

On fixe $\phi > 1$ vérifiant $p\phi < 1$. Nous venons de montrer que dans le disque centré en l'origine et de rayon ϕ et pour des motifs de taille suffisamment grande, le polynôme d'auto-corrélation ne s'annulait pas. Ce résultat va nous servir à satisfaire les hypothèses du théorème de Rouché (sur le cercle de rayon ϕ). Pour des tailles de motifs suffisamment grandes, la fonction $g(z)$ (ou plutôt $g_w(z)$, pour ne pas perdre de vue la dépendance en w) va être extrêmement faible (décroissance exponentielle) en module et que la fonction $f(z)$ va rester en module au-dessus d'une certaine constante positive.

La majoration $|g(z)| \leq (|z|p)^k$ est triviale et $|f(z)| \geq \alpha(\phi - 1)$. Puisque, par définition $p\phi < 1$, sur le cercle $\mathfrak{C}$ de rayon ϕ le module de la fonction g décroît exponentiellement, donc à partir d'un K_2 suffisamment grand

$$|g(z)| \leq (\phi p)^k < \alpha(\phi - 1) \leq |f(z)|.$$

On a donc satisfait les hypothèses du théorème de Rouché. Ainsi f et $f+g$ ont le même nombre de zéros dans le disque centré de rayon ϕ. Le lemme précédent montrait que dans le disque de rayon ϕ, le polynôme d'auto-corrélation ne s'annulait pas. Donc f ne s'annule que pour $z = 1$, *i.e.* il y a un unique zéro de f dans le disque centré de rayon ϕ. Le théorème de Rouché nous garantit donc que dans ce même disque il n'y a qu'une seule racine pour g. Cette racine est forcément ρ, la racine réelle dominante dont une valeur approchée a été trouvée dans la section précédente.

2.3.3 Détermination de la probabilité

Le théorème de Cauchy permet d'exprimer le coefficient d'ordre n de la série génératrice $\mathfrak{N}_w(z)$, $N_{n,w} = \mathbb{P}(\widehat{N}_w = 0)$ et le coefficient d'ordre n de la série génératrice $\mathfrak{O}_w(z)$, $O_{n,w} = \mathbb{P}(\widehat{N}_w = 1)$. Nous avons déjà obtenu une expression pour chacune de ces deux séries génératrices dans la section 2.2 ainsi que la localisation (et l'unicité) de leur unique pôle dominant ρ dans les sections 2.3.1 et 2.3.2. La notation est simplifiée en omettant l'indice w pour les coefficients N_n et O_n. Rappelons d'abord le théorème de Cauchy.

Théorème 3 (Cauchy) *Soit $\mathfrak{K}$ une courbe fermée simple orientée positivement dans une région ouverte simplement connexe Ω. Soit $g(z)$ une fonction méromorphe dans la région Ω et analytique*

sur la courbe $\mathfrak{K}$ et A l'ensemble des pôles de $g(z)$ à l'intérieur du contour ouvert défini par la courbe $\mathfrak{K}$. Alors

$$\frac{1}{2i\pi}\int_{\mathfrak{K}} g(s)ds = \sum_{s\in A}\mathrm{Res}(g,s).$$

Nous nous servons du théorème de Cauchy avec, comme contour le cercle $\mathfrak{C}$ centré en l'origine et de rayon ϕ déterminé précédemment dans le lemme 8. Les deux seuls pôles de la fonction $\mathfrak{O}(z)/z^{n+1}$ à l'intérieur du cercle $\mathfrak{C}$ sont 0 et ρ. De plus, la fonction $\mathfrak{O}(z)/z^{n+1}$ est analytique sur le cercle $\mathfrak{C}$. Ainsi :

$$I(\mathfrak{C}) := \frac{1}{2i\pi}\int_{\mathfrak{C}}\frac{\mathfrak{O}(z)}{z^{n+1}}dz = \mathrm{Res}\left(\frac{\mathfrak{O}(z)}{z^{n+1}};0\right) + \mathrm{Res}\left(\frac{\mathfrak{O}(z)}{z^{n+1}};\rho\right).$$

Le résidu en l'origine est le coefficient d'ordre n de la série génératrice $\mathfrak{O}(z)$. Nous cherchons à majorer $\mathfrak{O}(z)$ sur le contour $\mathfrak{C}$ pour contrôler la croissance de l'intégrale. Pour trouver une minoration du dénominateur $|c(z)(1-z)+p_w z^k|$ sur $|z|=\phi$, nous utilisons les résultats de la section précédente pour les motifs de taille supérieure à K :

$$\begin{aligned}|\mathfrak{D}(z)| \geq& |c(z)(1-z)| - |p_w z^k| > |c(z)||1-z| - (p\phi)^k\\ &\geq \alpha(\phi-1) - (p\phi)^k.\end{aligned}$$

Il existe un entier K_3 suffisamment grand tel que, pour tout $k \geq K_3$, $\alpha(\phi-1)-(p\phi)^k$ est plus grand qu'une constante positive κ. L'intégrale de $\mathfrak{O}(z)/z^{n+1}$ sur le contour est alors majorée par :

$$|I(\mathfrak{C})| \leq \frac{(p\phi)^k}{\kappa^2\phi^n} = O\left(\phi^{-n}\right).$$

La probabilité de n'avoir le motif w qu'à une seule reprise dans un texte de taille n s'écrit alors

$$O_n = [z^n]\mathfrak{O}(z) = -\mathrm{Res}\left(\frac{\mathfrak{O}(z)}{z^{n+1}};\rho\right) + O\left(\phi^{-n}\right),$$

pour un motif de taille supérieure à $\max\{K, K_3\}$.

Il nous faut maintenant calculer le résidu de $\mathfrak{O}(z)/z^{n+1}$ en ρ, pôle double de la série génératrice $\mathfrak{O}(z)$. Le dénominateur de $\mathfrak{O}(z)$ se développe au voisinage du pôle ρ (on a besoin du développement à l'ordre 3) et $z^{k-(n+1)}$ se développe au voisinage de ρ à l'ordre 1. Tous les termes en $\frac{1}{(z-\rho)}$ du développement en série de Laurent de la fonction $\mathfrak{O}(z)/z^{n+1}$ apparaissent. On arrive au résultat :

$$\mathrm{Res}\left(\frac{\mathfrak{O}(z)}{z^{n+1}};\rho\right) = \frac{p_w}{\mathfrak{D}'(\rho)^3}\rho^{k-n-2}((k-(n+1))\mathfrak{D}'(\rho) - \rho\mathfrak{D}''(\rho)) \tag{2.14}$$

où

$$\mathfrak{D}'(z) = kz^{k-1}p_w + (1-z)c'(z) - c(z),$$

et

$$\mathfrak{D}''(z) = k(k-1)z^{k-2}p_w + (1-z)c''(z) - 2c'(z).$$

La formule 2.14 ne donne pas une vision claire des termes dominants et des termes négligeables, une échelle va servir à simplifier ce résultat. L'échelle consiste en un développement de la formule selon les puissances de p_w. Le terme prépondérant de $((k-(n+1))\mathfrak{D}'(\rho)-\rho\mathfrak{D}''(\rho))/\mathfrak{D}'(\rho)^3$ va être le terme en $(p_w)^0 = 1$, c'est-à-dire $-(k-(n+1))c(\rho) + 2c'(\rho)$, mais puisque $k-(n+1)$

va être grand (n asymptotique) et ρ est proche de 1, le terme dominant est $nc(1)$. D'autre part le terme dominant dans $\mathfrak{D}'(\rho)^3$ est $-c^3(1)$ et $\rho^{k-2} = 1 + (k-2)p_w/c(1)$. On aboutit ainsi à

$$O_n = \frac{p_w}{c^2(1)}\rho^{-n}(n+1-k) + O\left(knp_w^2\right) + O\left(\phi^{-n}\right). \tag{2.15}$$

De plus le développement limité de ρ^{-n} s'écrit

$$\begin{aligned}\rho^{-n} &= \exp(-n\log\rho) = \exp\left(-n\left(\frac{p_w}{c(1)} + O(p_w^2)\right)\right)\\ &= \exp\left(-\frac{np_w}{c(1)}\right)\exp\left(-nO(p_w^2)\right) = \exp\left(-\frac{np_w}{c(1)}\right)(1 - nO(p_w^2)).\end{aligned}$$

Proposition 4 *La probabilité d'avoir exactement une occurrence du motif w parmi les n premières positions du texte engendré par une source sans mémoire a pour expression asymptotique*

$$\mathbb{P}_n(\widehat{N}_w = 1) = np_w\exp\left(-\frac{np_w}{c(1)}\right) + O\left(knp_w^2\right) + O(\phi^{-n}),$$

où les deux termes d'erreurs sont uniformes vis-à-vis de w.

Le raisonnement appliqué dans cette section pour le comportement asymptotique de la probabilité d'avoir exactement une seule occurrence du motif w est aussi valable (avec de très légères modifications) pour la probabilité de n'avoir aucune occurrence du motif w dans les n premières positions du texte.

Proposition 5 *La probabilité de n'avoir aucune occurrence du motif w parmi les n premières positions du texte engendré par une source sans mémoire vaut :*

$$\mathbb{P}_n(\widehat{N}_w = 0) = \exp\left(-\frac{np_w}{c(1)}\right) + O(\phi^{-n}),$$

où le terme d'erreur est uniforme vis-à-vis du motif w.

2.4 Étude de la différence pour la longueur de cheminement

Dans cette partie, nous montrons que les termes dominants du comportement asymptotique de l'espérance de la longueur de cheminement d'un trie et d'un arbre des suffixe sont identiques. Il est difficile de déterminer explicitement une expression asymptotique de la longueur de cheminement d'un arbre des suffixes. Le problème est contourné en montrant que la différence entre les deux longueurs de cheminement (pour un trie et un arbre des suffixes) est faible (en $O(n^{1-c})$ avec $c > 0$ dépendant des caractéristiques de la source, soit sous-linéaire) et les termes dominants du comportement asymptotique de la longueur de cheminement d'un arbre des suffixes seront ceux trouvés dans le chapitre 2 pour un trie. Nous obtenons le résultat suivant

Théorème 4 *L'espérance de la longueur de cheminement d'un arbre des suffixes construit sur les n premiers suffixes d'un mot produit par une source sans mémoire (p, q) avec $p \geq q$ et $0.5 \leq p \leq 0.54$ est donné quand n tend vers l'infini par :*

$$\mathbb{E}_n(L) = \frac{n\log n}{h} + \left(\frac{\gamma}{h} + \frac{p\log^2 p + q\log^2 q}{2h^2} + \epsilon(n)\right)n + O(n^{0.85}),$$

où $h = -p\log p - q\log q$ est l'entropie de la source, γ est la constante d'Euler[16] *et ϵ est une fonction fluctuant autour de zéro de très faible amplitude (typiquement inférieure à 10^{-5}).*

La moyenne de la longueur de cheminement d'un arbre des suffixes s'exprime en fonction de la probabilité $\mathbb{P}_n(\widehat{N}_w = 1)$ d'avoir une seule occurrence du motif w dans les n premières positions du texte et de la moyenne $\mathbb{E}_n(\widehat{N}_w)$ du nombre d'occurrences du motif w parmi les n premières positions d'un texte. Nous regardons la différence entre les comportements asymptotiques de l'espérance de la longueur de cheminement dans un trie sous modèle de Poisson de paramètre n et de l'espérance de la longueur de cheminement dans un arbre des suffixes construit sur les n premiers suffixes d'un texte. Asymptotiquement l'espérance du nombre d'occurrences du motif w dans les n premières positions du texte se comporte comme np_w, c'est-à-dire comme le comportement asymptotique de $\mathbb{E}_{\mathcal{P}(n)}(N_w)$. La différence entre l'espérance $\mathbb{E}^t_{\mathcal{P}(n)}(L)$ de la longueur de cheminement pour un trie sous modèle de Poisson de paramètre n et l'espérance $\mathbb{E}_n(L)$ de la longueur de cheminement pour un arbre des suffixes sous modèle de Bernoulli est approchée par

$$\begin{aligned}\mathbb{E}^t_{\mathcal{P}(n)}(L) - \mathbb{E}_n(L) &= \sum_{w\in\mathcal{A}^\star} \mathbb{E}(N_w) - \mathbb{P}(N_w = 1) - (\mathbb{E}(\widehat{N}_w) - \mathbb{P}(\widehat{N}_w = 1)) \\ &\simeq \sum_{w\in\mathcal{A}^\star} \mathbb{P}(\widehat{N}_w = 1) - \mathbb{P}(N_w = 1) \\ &\simeq \sum_{w\in\mathcal{A}^\star} np_w \left(\exp\left(-\frac{np_w}{c(1)}\right) - \exp(-np_w)\right) =: \Delta(n).\end{aligned} \tag{2.16}$$

La quantité $\Delta(n)$ est le terme dominant de la différence entre les deux moyennes (dans un trie et dans un arbre des suffixes).

Techniquement, on va décomposer la somme $\Delta(n)$ sur tous les motifs w en sommes sur différents ensembles définis selon la taille des motifs w : motifs courts, motifs intermédiaires et motifs longs. Les motifs courts sont peu nombreux et np_w tend vers l'infini. Cela nous permet de rendre leur contribution $\Delta_c(n)$ aussi petite que l'on veut. Pour les motifs longs, np_w tend vers zéro et cela nous permet d'utiliser un développement limité de la différence des deux exponentielles (les termes dominants sont les plus petites puissances). De plus la somme sur tous les motifs d'une même taille des carrés des probabilités décroît exponentiellement. Nous obtenons ainsi un comportement asymptotique sous-linéaire pour la somme $\Delta_l(n)$ sur les motifs longs. Les motifs intermédiaires sont divisés en deux sous-ensembles selon la valeur de leur polynôme d'auto-corrélation en 1. Il y a beaucoup de motifs dont la valeur en 1 du polynôme d'auto-corrélation est proche de 1, pour ceux-ci les deux exponentielles vont être proches même s'il faut réduire l'intervalle de validité de la probabilité. Peu de motifs intermédiaires ont une valeur de leur polynôme d'auto-corrélation « éloignée » de 1, cela nous permet de majorer leur contribution à la somme $\Delta(n)$.

Remarque : Puisque $c(1) \geq 1$, la somme $\Delta(n)$ est toujours une quantité positive.

2.4.1 Motifs « courts »

Dans cette section nous regardons la contribution des motifs « courts » à la somme $\Delta(n)$. On la note $\Delta_c(n)$. Il nous faut bien évidemment donner une définition précise de la « courteur » de ces motifs. Ensuite, la somme $\Delta_c(n)$ est majorée brutalement en se servant de la décroissance

[16] Leonhard Euler, mathématicien helvétique (1707–1783)

très rapide de la fonction $\exp(-x)$ vers 0 quand x tend vers l'infini et du nombre fini de motifs courts. Cette majoration nous garantit une contribution asymptotique négligeable de $\Delta_c(n)$ à la somme $\Delta(n)$. L'idée intuitive est que les motifs courts doivent avoir une probabilité pas trop petite, de façon à ce que np_w tende vers l'infini quand n tend vers l'infini.

Définition 7 *Un motif w est dit* court *pour la source (p, q) si*

$$|w| = k \leq \frac{5}{6} \log_{1/q} n = \frac{5}{6} C_q \log n =: k_c(n),$$

où $C_q := (-\log q)^{-1}$ et q est la plus petite des deux probabilités.

Ainsi pour n'importe quel motif court, il vient

$$np_w \geq nq^k = n \exp(k \log q) \geq n \exp\left(-\frac{5 \log q}{6 \log q} \log n\right) = n^{1/6}.$$

Le nombre total des motifs courts est donné par la somme des termes d'une progression géométrique : ce sont tous les motifs de taille inférieure à $(5/6)C_q \log n$. Il y en a de l'ordre de $2^{(5/6) \lg n} = n^{5/6}$ pour une source symétrique (où lg est le logarithme de base 2) et $n^{5C_q/6}$ dans le cas biaisé.

La somme $\Delta_c(n)$ est définie par

$$\Delta_c(n) := \sum_{k=0}^{k_c(n)} \sum_{w \in \mathcal{A}^k} np_w \left(\exp\left(-\frac{np_w}{c_w(1)}\right) - \exp(-np_w) \right).$$

La somme $\Delta_c(n)$ est majorée brutalement par le produit du plus grand terme individuel et du nombre de motifs courts :

$$\Delta_c(n) \;\leq \sum_{w:\, |w| \leq k_c(n)} np_w \exp\left(-\frac{np_w}{c(1)}\right) \;\leq\; \#\{w : |w| \leq k_c(n)\} \max_{w \text{ court}} \left\{ np_w \exp\left(-\frac{np_w}{c(1)}\right) \right\}.$$

Le cardinal de l'ensemble des motifs courts a déjà été déterminé ; d'autre part, la fonction $x \to x \exp(-x/c(1))$ est décroissante pour x suffisamment grand et donc son maximum sur les motifs courts est obtenu pour les plus petites valeurs de np_w soit $n^{1/6}$. De plus, $c(1) \leq (1-p)^{-1} =: K_p^{-1}$ donc

$$\Delta_c \;\leq\; n^{\frac{5}{6}C_q} n^{\frac{1}{6}} \exp\left(-\frac{n^{1/6}}{c(1)}\right) \;\leq\; n^{\frac{1+5C_q}{6}} \exp\left(-n^{1/6} K_p\right). \tag{2.17}$$

Les deux constantes C_q et K_p ne jouent aucun rôle dans le comportement asymptotique de $\Delta_c(n)$ et la décroissance très rapide de la fonction exponentielle permet d'obtenir le

Lemme 9 *La contribution $\Delta_c(n)$ des motifs courts (de taille inférieure à $k_c(n)$) à la somme $\Delta(n)$ est exponentiellement décroissante. Pour la suite de ce chapitre, on écrit seulement*

$$\Delta_c(n) = o(1). \tag{2.18}$$

La formulation de l'équation (2.18) nous suffit pour la suite de l'étude mais on peut évidemment l'améliorer puisque la décroissance est *de facto* exponentielle.

Remarque : La définition des motifs courts est relativement arbitraire. Les résultats auraient été les mêmes avec une borne $k_c(n) = (2/3)C_q \log n$ par exemple. En fait, toutes les bornes $k_c(n)$ du type $\alpha C_q \log n$ avec $\alpha < 1$ sont valables et permettent d'obtenir une décroissance exponentielle de $\Delta_c(n)$ avec n et donc un comportement asymptotique négligeable.

2.4.2 Motifs « longs »

Cette partie traite des motifs « longs », c'est-à-dire les motifs dont la taille est supérieure à une borne $k_l(n)$. Leur contribution à la somme $\Delta(n)$ est notée $\Delta_l(n)$. Ces motifs longs sont en nombre infini et donc la majoration grossière de la sous-section précédente ne peut fonctionner. Leur taille va croître suffisamment vite avec n pour que np_w tende vers zéro. Un développement limité du terme général de la somme nous en fait ressortir le terme dominant. La propriété de décroissance exponentielle des probabilités de coïncidences $\sum_{w\in\mathcal{A}^k} p_w^2$ permet de conclure.

Définition 8 *Un motif w est dit* long *pour la source (p,q) si*

$$|w| = k \geq \frac{3}{2}\log_{1/p} n = 1.5 C_p \log n =: k_l(n),$$

où p est la plus grande des deux probabilités et $C_p = (-\log p)^{-1}$.

Ainsi pour n'importe quel motif long de taille k,

$$np_w \leq np^k = n\exp(k\log p) \leq \frac{1}{\sqrt{n}}$$

et la quantité np_w tend vers zéro quand n tend vers l'infini. Les premiers termes du développement limité de la fonction exponentielle sont donc les termes dominants, soit

$$\exp\left(-\frac{np_w}{c(1)}\right) - \exp(-np_w) \sim np_w\left(1-\frac{1}{c(1)}\right).$$

Tous les termes de la somme sont positifs puisque $c(1)\geq 1$ donc la contribution $\Delta_l(n)$ des motifs longs à la somme Δ est approchée, quand n tend vers l'infini, par

$$\Delta_l(n) \simeq \sum_{k\geq k_l(n)} \sum_{w\in\mathcal{A}^k} (np_w)^2\left(1-\frac{1}{c(1)}\right).$$

La somme $\sum_{w\in\mathcal{A}^k}(p_w)^2$ possède une propriété générale appelée décroissance exponentielle des probabilités de coïncidence. Dans le cas d'une source symétrique, $p_w = 2^{-k}$, quantité indépendante de w, et la somme vaut 2^{-k}. Si la source est biaisée (p,q), on a :

$$\sum_{w\in\mathcal{A}^k}(p_w)^2 \;=\; \sum_{i=0}^{k}\binom{k}{i}\left(p^i q^{k-i}\right)^2 \;=\; \sum_{i=0}^{k}\binom{k}{i}\left(p^2\right)^i\left(q^2\right)^{k-i} \;=\; \left(p^2+q^2\right)^k$$

en utilisant la formule du binôme de Newton.

Dans les deux cas (en fait le cas symétrique n'est que la particularisation à $p=q=1/2$ de la source (p,q)), il existe une constante $A_p := p^2+q^2$ plus petite que 1, dépendante de p, telle que

$$\sum_{w\in\mathcal{A}^k}(p_w)^2 = A_p^k.$$

Ce résultat est utilisé dans le développement limité de la somme $\Delta_l(n)$:

$$\sum_{k\geq k_l(n)}\sum_{w\in\mathcal{A}^k}(np_w)^2\left(1-\frac{1}{c(1)}\right) = O\left(\sum_{k\geq k_l(n)} n^2 A_p^k\right).$$

La valeur en 1 du polynôme d'auto-corrélation probabilisé vérifie $1 \leq c(1) \leq 1/(1-p)$. La notation O est donc justifiée par la majoration de $1 - 1/c(1)$ par p.

En sommant sur toutes les tailles supérieures à $k_l(n)$, le comportement asymptotique de la somme $\Delta_c(n)$ est en $O\left(n^2 A_p^{k_l(n)}\right)$. La quantité $p^2 + q^2 = 1 - 2p + 2p^2$ est toujours inférieure à p pour $p \in [1/2, 1[$, on obtient donc

Lemme 10 *La somme $\Delta_l(n)$ se comporte asymptotiquement en*

$$\Delta_l(n) = \sum_{k \geq k_l(n)} \sum_{w \in \mathcal{A}^k} np_w \left(\exp\left(-\frac{np_w}{c_w(1)} \right) - \exp(-np_w) \right) = O\left(\sqrt{n}\right).$$

Remarque : Comme dans la section sur les motifs courts, la borne $k_l(n)$ est assez arbitraire. La méthode s'applique à n'importe quelle borne $k_l(n) = \beta C_p \log n$ pour $\beta > 1$. L'exposant du comportement asymptotique devient $2 - \beta$.

2.4.3 Motifs « intermédiaires »

Cette partie traite des « autres » motifs, ceux qui ne rentrent dans aucune des deux parties précédentes. Ces motifs sont nommés « intermédiaires. » Leur taille varie entre

$$k_c(n) = (5/6)C_q \log n \text{ et } k_l(n) = 1.5C_p \log n.$$

Nous montrons que la contribution de ces motifs est sous-linéaire. Pour la plupart des motifs intermédiaires w, la valeur en 1 du polynôme d'auto-corrélation est très proche de 1. La différence entre $\exp(-np_w/c(1))$ et $\exp(-np_w)$ va donc être faible. Ces motifs sont appelés *apériodiques* intermédiaires. Néanmoins, il existe des motifs pour lesquels $c_w(1)$ n'est pas suffisamment proche de 1 pour que la différence soit faible. Le nombre de ces motifs, nommés *périodiques* intermédiaires, est extrêmement limité et cela nous permet de contrôler leur contribution asymptotique à la somme $\Delta(n)$.

Définition 9 *Soit une source sans mémoire fixée, un motif w de taille k est dit* périodique *pour cette source si la valeur de son polynôme d'auto-corrélation probabilisé en 1 vérifie*

$$c(1) \geq 1 + 2^{-k/2}.$$

Un motif est dit apériodique *sinon. L'ensemble des motifs périodiques de taille k est*

$$\mathcal{B}_k := \{w : |w| = k, \quad c(1) \geq 1 + 2^{-k/2}\}.$$

➜ Motifs périodiques intermédiaires

Nous traitons d'abord le cas des motifs intermédiaires et périodiques. Pour ces motifs, la valeur du polynôme d'auto-corrélation en 1 est assez éloignée de 1, cependant le nombre de ces motifs est faible et ainsi leur contribution que l'on note $\Delta_p(n)$ est asymptotiquement assez faible.

On va introduire un lemme qui permet de minorer le nombre de motifs périodiques intermédiaires :

Lemme 11 *Le nombre de motifs périodiques de taille k est borné par $k2^{k/2}$, soit*

$$\#\mathcal{B}_k < k2^{k/2}. \tag{2.19}$$

Preuve : Le lemme 2.10 de la page 43 donne une expression de la somme des valeurs du polynôme d'auto-corrélation en 1 sur tous les motifs d'une taille donnée, et ce quel que soit la valeur des probabilités :

$$\sum_{w\in\mathcal{A}^k} c(1) = 2^k + k - 1.$$

Cette somme sur tous les motifs de taille k se scinde en, d'un côté, la somme sur les motifs qui appartiennent à $\mathcal{B}_k$ et, de l'autre côté, la somme sur les motifs apériodiques :

$$\sum_{w\in\mathcal{A}^k} c(1) = \sum_{w\in\mathcal{B}_k} c(1) + \sum_{w\notin\mathcal{B}_k} c(1).$$

Si w appartient à $\mathcal{B}_k$, on a par définition $c(1) \geq 1+2^{-k/2}$. Ainsi la contribution des motifs de $\mathcal{B}_k$ est minorée par $\#\mathcal{B}_k(1+2^{-k/2})$. Pour les autres motifs, on minore la valeur en 1 du polynôme d'auto-corrélation par 1. On trouve

$$2^k + k - 1 \geq \#\mathcal{B}_k(1+2^{-k/2}) + (2^k - \#\mathcal{B}_k),$$

et la résolution de cette inéquation en fonction du cardinal de l'ensemble $\mathcal{B}_k$ donne le résultat souhaité. ◀

À partir de ce lemme, il est facile de majorer la contribution $\Delta_p(n)$ des motifs périodiques intermédiaires à la somme $\Delta(n)$ par

$$\begin{aligned}
\Delta_p(n) &:= \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{B}_k} np_w \left(\exp\left(-\frac{np_w}{c(1)}\right) - \exp(-np_w)\right) \\
&\leq \sum_{k=k_c(n)}^{k_l(n)} \#\mathcal{B}_k\, np^k \max_{w\in\mathcal{B}_k}\left\{\exp\left(-\frac{np_w}{c(1)}\right) - \exp(-np_w)\right\} \\
&\leq Kn \sum_{k=k_c(n)}^{k_l(n)} k2^{k/2}p^k \;\leq\; K'nk_c(n)(p\sqrt{2})^{k_c(n)} \\
&\leq K''nn^{(5/6)C_q\log(p\sqrt{2})}\log n,
\end{aligned} \tag{2.20}$$

où K est le maximum de la fonction $x \to \exp(-x/c(1)) - \exp(-x)$ sur les réels positifs, K' et K'' deux autres constantes. Ce maximum est fini puisque la fonction est continue et tend vers zéro en l'infini. De plus il faut que le terme général $k(p\sqrt{2})^k$ de la somme tende vers zéro, c'est-à-dire $p\sqrt{2} < 1$.

Finalement on arrive à une contribution asymptotique des motifs intermédiaires périodiques en

$$\Delta_p(n) = O\left(n^{1+(5/6)C_q\log(p\sqrt{2})}\log n\right).$$

Notre but est d'avoir une contribution sous-linéaire de $\Delta_p(n)$, il faut que

$$\frac{5}{6}C_q\log(p\sqrt{2}) = -\frac{5\log(p\sqrt{2})}{6\log q} < 0,$$

ce qui signifie encore que $p\sqrt{2} < 1$.

Lemme 12 *La somme $\Delta_p(n)$ se comporte asymptotiquement en*

$$O\left(n^{1+(5/6)C_q \log(p\sqrt{2})} \log n\right)$$

quand la borne des motifs courts est $k_c(n) = (5/6)C_q \log n$. Si on souhaite un comportement asymptotique sous-linéaire de la somme $\Delta_p(n)$, la probabilité p doit se situer dans l'intervalle $[0.5, \frac{1}{\sqrt{2}}[$.

➜ Motifs apériodiques intermédiaires

Les motifs *apériodiques* intermédiaires sont ceux dont la taille est intermédiaire et qui n'appartiennent pas à l'ensemble $\mathcal{B}_k$. La contribution des motifs apériodiques à la somme $\Delta(n)$ est notée $\Delta_a(n)$. Par définition, la valeur en 1 du polynôme d'auto-corrélation de ces motifs est majorée par $1 + 2^{-k/2}$. Dans cette section la taille des motifs est intermédiaire, c'est-à-dire suffisamment grande pour que la valeur de $2^{-k/2}$ soit assez proche de zéro. La différence des deux exponentielles $\exp(-np_w/c(1))$ et $\exp(-np_w)$ dans notre formule va rester faible. Ainsi, même s'il y a beaucoup de motifs apériodiques-intermédiaires (de l'ordre de 2^k pour chaque taille k), leur contribution individuelle va être suffisamment faible pour que la somme soit asymptotiquement en $O(n^{1-c})$.

Les motifs apériodiques-intermédiaires vérifient par définition

$$\frac{1}{c(1)} \geq \frac{1}{1+2^{-k/2}} \geq 1 - 2^{-k/2},$$

et, en se servant de la croissance de la fonction exponentielle, le terme général est

$$np_w \left(\exp\left(-\frac{np_w}{c(1)}\right) - \exp(-np_w)\right) \leq np_w \exp(-np_w)\left(\exp\left(np_w 2^{-k/2}\right) - 1\right).$$

Les termes dans les deux fonctions exponentielles sont proches, l'utilisation d'un développement limité de la fonction exponentielle est naturelle. Pour que les termes qui priment dans la développement de $\exp(np_w 2^{-k/2})$ soient les petites puissances, il faut que $np_w 2^{-k/2}$ tende vers zéro. Quitte à contraindre la probabilité p, les motifs apériodiques vérifient $np_w 2^{-k/2} \to 0$. Soit, pour toute longueur intermédiaire k, p satisfait

$$np^k 2^{-k/2} \to 0. \tag{2.21}$$

Cela contraint la borne supérieure des motifs courts $k_c(n)$ à vérifier

$$\frac{5}{6} C_q \log\left(\frac{p}{\sqrt{2}}\right) + 1 < 0,$$

soit

$$\left(\frac{p}{\sqrt{2}}\right)^{5/6} + p - 1 < 0. \tag{2.22}$$

Après calcul numérique, on trouve la condition

$$p < p_1 \simeq 0.5469205467,$$

où p_1 est l'unique solution de (2.22) dans $[0,1]$. Une fois la condition sur la probabilité vérifiée, le développement limité de la fonction exponentielle au voisinage de zéro fait ressortir un terme dominant

$$(np_w)^2 \exp(-np_w) 2^{-k/2},$$

pour le terme général de la somme $\Delta_a(n)$. La fonction $x \to x^2 \exp(-x)$ est bornée par $\delta = 4e^{-2}$ (pour x positif), la contribution générale de ces motifs devient

$$\Delta_a(n) \leq \sum_{k=k_c(n)}^{k_l(n)} \delta 2^k 2^{-k/2} = O\left(n^{(3/4)C_p \log(2)}\right).$$

Le choix de la borne inférieure de l'ensemble des motifs intermédiaires (qui est la borne supérieure des motifs courts) détermine l'intervalle de validité sur p du comportement asymptotique de $\Delta_a(n)$. L'exposant du comportement asymptotique dépend quant à lui de la borne supérieure $k_l(n)$. L'intervalle de validité sur p du comportement asymptotique ne peut s'agrandir qu'au dépens de la petitesse du terme d'erreur. L'exposant est inférieur à 1 pour $p < p_0 = 2^{-3/4} \simeq 0.5946035575$.

Lemme 13 *La contribution asymptotique de la somme des motifs apériodiques intermédiaires* $\Delta_a(n)$ *est en*

$$O\left(n^{(3/4)C_p \log 2}\right),$$

pour une probabilité $p < 0.54$ *et une borne supérieure des motifs intermédiaires* $k_l(n) = 1.5 C_p \log n$.

Remarque : En modifiant le coefficient ($\alpha = 5/6$) de la borne $k_c(n)$ (*i.e.* en le rapprochant de 1), on peut atteindre une borne supérieure sur la probabilité p de $2 - \sqrt{2} \simeq 0.5857864376$.

➜ Pourquoi « périodiques » ?

Pour comprendre la genèse de l'appellation « périodique » (ou apériodique) des motifs, il faut revenir au cas d'une source symétrique.

Le coefficient d'indice 0 du polynôme d'auto-corrélation vaut toujours 1. Soit i le plus petit indice non nul pour lequel le coefficient c_i du polynôme d'auto-corrélation est non-nul. Ce coefficient peut ne pas exister si la seule corrélation dans le motif w est la corrélation triviale du motif avec lui-même, alors le polynôme d'auto-corrélation vaut 1 et i vaut la taille k du motif. Nous présentons la définition originelle des motifs périodiques, elle n'est valable **que** dans cette sous-section :
Un motif w de taille k est dit périodique *si le premier indice strictement positif d'un coefficient non nul du polynôme d'auto-corrélation est situé avant $k/2$. Soit si*

$$\min_{i>0}\{c_i \neq 0\} \leq \frac{k}{2}.$$

Un motif est dit apériodique *sinon.*
Remarque : Contrairement à la définition 9, la définition des motifs périodiques et apériodiques ci-dessus ne tient pas compte du modèle probabiliste, juste de la structure interne du motif.

Soit w un motif périodique, le premier 1 non trivial dans l'écriture de son polynôme d'auto-corrélation est situé avant $k/2$ donc la valeur en 1 de leur polynôme d'auto-corrélation vaut au minimum $1 + 2^{-k/2}$. Il est donc naturel d'introduire l'ensemble

$$\mathcal{B}_k := \{w : |w| = k, \quad c(1) \geq 1 + 2^{-k/2}\}, \tag{2.23}$$

qui est l'ensemble des motifs périodiques au sens.

Si le polynôme d'auto-corrélation a son premier indice non-nul i avant $k/2$ alors le suffixe de taille i va se répéter dans le motif w comme on l'a déjà vu dans la preuve du lemme 2.10. Le terme de « périodique » vient de cette répétition du suffixe de taille i.

Pour le cas d'une source biaisée, on garde exactement le même ensemble $\mathcal{B}_k$ car l'application de la notion de périodicité *stricto sensu* n'est pas facilement manipulable. La valeur du polynôme d'auto-corrélation ne dépend plus que de la structure combinatoire du motif w mais aussi des lettres qui le compose (et les deux lettres `0` et `1` n'ont plus la même probabilité).

2.5 Étude de la différence pour la taille

Dans cette section, nous étudions la différence entre la moyenne de la taille d'un trie sous modèle de Poisson de paramètre n et la moyenne de la taille d'un arbre des suffixes construit en prenant les n premiers suffixes d'un texte engendrés par une source sans mémoire. L'expression de la moyenne de la taille d'un trie, ainsi que son comportement asymptotique ont été rappelés dans le chapitre précédent. D'autre part, les deux probabilités qui entrent dans l'écriture de la taille d'un arbre des suffixes viennent d'être déterminées asymptotiquement dans la section 2.3.3. Nous regardons le comportement asymptotique de la différence en découpant, comme pour la longueur de cheminement, l'ensemble des motifs en motifs longs périodiques, longs apériodiques, courts, intermédiaires périodiques et intermédiaires apériodiques. Les méthodes ne sont pas exactement identiques mais en de nombreux points similaires à celles de la section 2.4 ce qui nous permet d'alléger les calculs. Nous montrons que

Théorème 5 *L'espérance de la taille d'un arbre des suffixes construit sur les n premiers suffixes d'un mot produit par une source sans mémoire (p, q) avec $p \geq q$ et $0.5 \leq p \leq 0.54$ est donné quand n tend vers l'infini par :*

$$\mathbb{E}_n(S) = \frac{n}{h}(1 + \epsilon'(n)) + O(n^{0.85}),$$

où $h = -p \log p - q \log q$ est l'entropie de la source et ϵ' est une fonction oscillant autour de zéro de très faible amplitude (inférieure à 10^{-5}).

La taille a été définie par l'équation 1.1 de la page 20 par

$$S = \sum_{w \in \mathcal{A}^\star} [\![N_w \geq 2]\!].$$

L'expression de la moyenne de la taille d'un arbre des suffixes est

$$\mathbb{E}_n(S) = \sum_{w \in \mathcal{A}^\star} 1 - \mathbb{P}_n(\widehat{N}_w = 0) - \mathbb{P}_n(\widehat{N}_w = 1). \tag{2.24}$$

Le comportement asymptotique de la différence entre la taille moyenne $\mathbb{E}^t_{\mathcal{P}(n)}(S)$ d'un trie où le nombre de chaînes suit une loi de Poisson de paramètre n et la taille moyenne $\mathbb{E}_n(S)$ d'un arbre des suffixes construit sur les n premiers suffixes d'un texte est approchée par

$$\begin{aligned}\mathbb{E}^t_{\mathcal{P}(n)}(S) - \mathbb{E}_n(S) &= \sum_{w \in \mathcal{A}^\star} \mathbb{P}(\widehat{N}_w = 0) + \mathbb{P}(\widehat{N}_w = 1) - \mathbb{P}(N_w = 0) - \mathbb{P}(N_w = 1) \\ &\simeq \sum_{w \in \mathcal{A}^\star} (1 + np_w) \left(\exp\left(-\frac{np_w}{c(1)} \right) - \exp(-np_w) \right) =: \mathcal{D}(n).\end{aligned}$$

La quantité $\mathcal{D}(n)$ est le terme dominant de la différence entre les deux moyennes (dans un trie et dans un arbre des suffixes). L'expression des probabilités pour le trie est une simple application de la loi de Poisson. Dans le cas de l'arbre des suffixes, les probabilités $\mathbb{P}(\widehat{N}_w = 0)$ et $\mathbb{P}(\widehat{N}_w = 1)$ viennent d'être déterminées asymptotiquement dans la section 2.3.3.

2.5.1 Motifs courts

Les motifs courts sont les motifs de taille inférieure à

$$k_c(n) := (5/6)C_q \log n \text{ où } C_q := (-\log q)^{-1}.$$

La contribution des motifs courts à la somme $\mathcal{D}(n)$ est notée $\mathcal{D}_c(n)$. Pour ces motifs, la quantité np_w tend vers l'infini avec n. Le nombre de motifs courts est fini et de l'ordre de $n^{5C_q/6}$. La majoration brutale a prouvé son efficacité dans le cas de la longueur de cheminement, elle est donc reprise ici :

$$\begin{aligned}\mathcal{D}_c(n) &:= \sum_{k=0}^{k_c(n)} \sum_{w\in\mathcal{A}^k} (1+np_w)\left(\exp\left(-\frac{np_w}{c(1)}\right) - \exp(-np_w)\right)\\ &\leq \#\{w : |w| \leq k_c(n)\} \max_{w \text{ court}} \left\{(1+np_w)\exp\left(-\frac{np_w}{c(1)}\right)\right\}.\end{aligned}$$

La fonction $x \mapsto (1+x)\exp(-x/c(1))$ est décroissante pour x au voisinage de l'infini, donc elle prend sa plus grande valeur pour les plus petites valeurs de x. La plus petite valeur de np_w est $n^{1/6}$, ainsi la quantité $\mathcal{D}_c(n)$ est bornée par

$$\mathcal{D}_c(n) \leq n^{5C_q/6}\left(1+n^{1/6}\right)\exp\left(-n^{1/6}K_p\right),$$

où $K_p := (1-p) > 0$.

À partir de ce résultat nous obtenons :

Lemme 14 *La contribution $\mathcal{D}_c(n)$ des motifs courts, c'est-à-dire des motifs de taille inférieure à $k_c(n)$, est en $o(1)$ quand n tend vers l'infini.*

Remarque 1 : La borne définissant les motifs courts est, comme pour la longueur de cheminement, arbitraire et le lemme ci-dessus est valable pour toute borne supérieure $k_c(n) = \alpha C_q \log n$, avec $\alpha < 1$.
Remarque 2 : La contribution est en fait exponentiellement décroissante, mais la formulation du lemme est suffisante pour la suite.

2.5.2 Motifs longs

Les motifs longs sont définis comme les motifs dont la taille est supérieure à $k_l(n)$. La borne inférieure $k_l(n)$ est la même que dans la section 2.4.1 sur la longueur de cheminement, c'est-à-dire

$$k_l(n) = 1.5C_p \log n.$$

La contribution des motifs longs à la somme $\mathcal{D}(n)$ est notée $\mathcal{D}_l(n)$. La quantité np_w tend vers zéro quand n tend vers l'infini. Le terme général

$$\Psi_w(n) := (1+np_w)\left(\exp\left(-\frac{np_w}{c(1)}\right) - \exp(-np_w)\right)$$

de la différence $\mathcal{D}(n)$ est aussi une fonction de np_w. Son développement limité au voisinage de zéro donne un terme dominant en

$$np_w\left(1-\frac{1}{c(1)}\right).$$

La méthode utilisée pour la longueur de cheminement (qui faisait appel à la propriété de décroissance exponentielle des probabilités de coïncidence) ne peut plus fonctionner puisqu'on a juste np_w et non plus np_w^2. Il faut séparer les motifs longs en deux sous-ensembles : les motifs longs périodiques et longs apériodiques. Les définitions de ces deux ensembles sont les mêmes que dans la section 2.4.3 : les motifs périodiques sont ceux dont la valeur en 1 du polynôme d'auto-corrélation est plus grande que $1+2^{-k/2}$.

➜ Motifs longs périodiques.

La contribution des motifs longs **et** périodiques à la somme $\mathcal{D}(n)$ est notée $\mathcal{D}_{l,p}(n)$. Rappelons que $\mathcal{B}_k$ est l'ensemble des motifs périodiques de taille k. Pour les motifs longs, la quantité np_w tend vers zéro. En effectuant un développement limité du terme général Ψ_w au voisinage de l'origine, il vient

$$\begin{aligned}\mathcal{D}_{l,p}(n) &:= \sum_{k\geq k_l(n)}\sum_{w\in\mathcal{B}_k}(1+np_w)\left(\exp\left(-\frac{np_w}{c(1)}\right)-\exp\left(-np_w\right)\right)\\ &\simeq \sum_{k\geq k_l(n)}\sum_{w\in\mathcal{B}_k}(1+np_w)np_w\left(1-\frac{1}{c(1)}\right).\end{aligned}$$

La majoration $c_w(1)\leq(1-p)^{-1}$, où p est la plus grande des deux probabilités, est valable pour n'importe quel motif. Il en résulte la majoration

$$1-\frac{1}{c(1)}\leq p.$$

De plus, le nombre de motifs périodiques de taille k est borné (lemme 11) par $k2^{k/2}$. Ainsi, en majorant brutalement la contribution des motifs longs périodiques par le produit du nombre de motifs périodiques et de leur contribution maximale, on obtient

$$\mathcal{D}_{l,p}(n)\leq\sum_{k\geq k_l(n)}k2^{k/2}np_w\left(1-\frac{1}{c(1)}\right)\leq pn\sum_{k\geq k_l(n)}k(p\sqrt{2})^k.$$

Pour que la somme des $k(p\sqrt{2})^k$ converge il faut que $p\sqrt{2}<1$, c'est-à-dire $p<1/\sqrt{2}$. Avec cette condition, la contribution des motifs longs et périodiques est majorée par

$$\mathcal{D}_{l,p}(n)\leq pn\frac{\left(1.5C_p\log n(1-p\sqrt{2})+p\sqrt{2}\right)}{(1-p\sqrt{2})^2}(p\sqrt{2})^{1.5C_p\log n}=O(n^\alpha\log n),$$

où l'exposant α vaut $1-1.5\log(p\sqrt{2})/\log p$. Pour $p<\sqrt{2}$, cet exposant est plus petit que 1 ce qui nous offre bien un comportement asymptotique sous-linéaire de la somme $\mathcal{D}_{l,p}(n)$.

➜ **Motifs longs apériodiques.**

La somme sur les motifs longs **et** apériodiques est notée $\mathcal{D}_{l,a}(n)$. Deux arguments vont jouer dans la majoration de la contribution des motifs longs et apériodiques. D'une part la borne sur les valeurs que prend le polynôme d'auto-corrélation en 1 (borne liée à la définition de l'ensemble $\mathcal{B}_k$) permet de majorer $1 - c^{-1}(1)$. D'autre part, on développe la fonction $\exp(np_w)$ pour np_w proche de zéro.

Puisque les motifs sont apériodiques, la valeur en 1 de chaque polynôme d'auto-corrélation vérifie

$$\frac{1}{c_w(1)} \geq \frac{1}{1+2^{-k/2}} \geq 1 - 2^{k/2},$$

où k est la taille du motif w. La croissance de la fonction exponentielle nous apporte la majoration

$$(1+np_w)\left(\exp\left(-\frac{np_w}{c(1)}\right) - \exp\left(-np_w\right)\right) \leq B\exp(-np_w)\left(\exp\left(np_w 2^{-k/2}\right) - 1\right), \qquad (2.25)$$

pour une constante $B > 1$.

L'avantage d'un développement limité de l'exponentielle au voisinage de zéro, c'est que le terme dominant est celui de plus petite puissance. Pour les motifs de taille supérieure à $k_l(n)$, $np_w 2^{-k/2}$ tend vers 0 pour toute probabilité p et on développe $\exp(np_w 2^{-k/2})$ au voisinage de zéro.

Le nombre de motifs apériodiques de taille k est trivialement borné par 2^k. Puisque tous les termes dans la somme sont positifs, la somme $\mathcal{D}_{l,a}(n)$ est bornée par

$$\sum_{k\geq k_l(n)} \sum_{w\notin\mathcal{B}_k} \exp(-np_w)\left(np_w 2^{-k/2}\right) \leq \sum_{k\geq k_l(n)} 2^k np_w 2^{-k/2} \leq n \sum_{k\geq k_l(n)} (p\sqrt{2})^k.$$

Il faut impérativement que la raison $p\sqrt{2}$ de la progression géométrique soit inférieure à 1 pour que la somme converge. Cela force la condition $p < (\sqrt{2})^{-1}$ et le comportement asymptotique est alors

$$\mathcal{D}_{l,a}(n) = O\left(n(p\sqrt{2})^{k_l(n)}\right).$$

L'exposant du comportement asymptotique de $\mathcal{D}_{l,a}(n)$ est $1 + 1.5C_p\log(p\sqrt{2})$ et il est inférieur à 1 pour $p < (\sqrt{2})^{-1}$.

En assemblant les résultats pour les motifs longs périodiques et apériodiques, il vient

Lemme 15 *La contribution des motifs longs (de taille supérieure à $k_l(n) = 1.5C_p\log n$) à la somme $\mathcal{D}(n)$ est de l'ordre de*

$$\mathcal{D}_l(n) = O\left(n^{1+1.5C_p\log(p\sqrt{2})}\log n\right),$$

pour $p\sqrt{2} < 1$ et n tendant vers l'infini.

Remarque : En contraignant un peu plus le domaine de validité de la probabilité, on obtient une majoration plus fine de la croissance du terme $\mathcal{D}_l(n)$.

2.5.3 Motifs intermédiaires.

Les motifs intermédiaires sont les motifs dont la taille varie entre

$$k_c(n) = (5/6)C_q \log n \text{ et } k_l(n) = 1.5C_p \log n, \text{ avec } C_r = (\log r)^{-1}.$$

Comme dans le cas de la longueur de cheminement, l'ensemble des motifs intermédiaires est scindé en deux : les motifs périodiques et apériodiques.

Pour les motifs périodiques intermédiaires, la méthode est exactement la même que dans la sous-section 2.4.3. Il y a au plus $k2^{k/2}$ motifs périodiques de taille k. La contribution des motifs périodiques intermédiaires est majorée comme dans l'équation (2.20) et le comportement asymptotique pour la contribution des motifs intermédiaires et périodiques est de l'ordre de $O(n^{(3/4)C_p \log 2} \log n)$ pour $p < 2^{-3/4} = p_0 \simeq 0.5946035575$.

La contribution des motifs intermédiaires apériodiques à la somme $\mathcal{D}(n)$ est notée $\mathcal{D}_{i,a}(n)$Les motifs apériodiques intermédiaires sont traités de manière similaire au paragraphe 2.4.3. Pourvu que $np^k 2^{-k/2}$ tende vers zéro (c'est-à-dire $p < p_1 \simeq 0.547$), le majoration de l'équation (2.25) permet d'obtenir

$$\begin{aligned} \mathcal{D}_{i,a}(n) &:= \sum_{k_c(n)}^{k_l(n)} \sum_{w \notin \mathcal{B}_k} (1 + np_w) \left(\exp\left(-\frac{np_w}{c(1)} \right) - \exp\left(-np_w \right) \right) \\ &\leq \sum_{k_c(n)}^{k_l(n)} 2^k (1 + np_w) \exp(-np_w) np_w 2^{-k/2} \leq C n^{(3/4)C_p \log 2}, \end{aligned} \tag{2.26}$$

pour une constante C. La quantité $(3/4)C_p \log 2$ est inférieure à 1 pour $p < p_0$.

Lemme 16 *La contribution asymptotique de la somme des motifs intermédiaires est en*

$$O\left(n^{(3/4)C_p \log 2} \log n \right),$$

pour $p < 0.54$ et n tendant vers l'infini.

2.6 Conclusion

Dans la section 2.4, les différentes contributions asymptotiques des motifs courts, longs, intermédiaires périodiques et intermédiaires apériodiques à la somme $\Delta(n)$ ont été obtenues. Ces résultats ne sont valables que pour $0.5 \leq p \leq 0.54$ (cette borne vient des motifs apériodiques intermédiaires). Avec cette (forte) contrainte sur la probabilité p, pour une borne supérieure des motifs courts $k_c(n) = (5/6)C_q \log n$ et une borne inférieure des motifs longs $k_l(n) = 1.5C_p \log n$, la différence $\Delta(n)$ se comporte asymptotiquement en $O(n^{0.85})$. Nous avons retrouvé le comportement asymptotique de la moyenne de la longueur de cheminement d'un trie au chapitre précédent, il vient donc :

Théorème (longueur de cheminement) *Le comportement asymptotique de l'espérance de la longueur de cheminement d'un arbre des suffixes construit sur les n premiers suffixes d'un mot produit par une source sans mémoire (p, q) avec $p \geq q$ et $0.5 \leq p \leq 0.54$ est donné par :*

$$\mathbb{E}_n(L) = \frac{n \log n}{h} + \left(\frac{\gamma}{h} + \frac{p \log^2 p + q \log^2 q}{2h^2} + \epsilon(n) \right) n + O(n^{0.85}),$$

où $h = -p\log p - q\log q$ *est l'entropie de la source,* γ *est la constante d'Euler et* ϵ *est une fonction fluctuante de très faible amplitude (typiquement inférieure à* 10^{-5}*).*

Dans la section 2.5 les différentes contributions à la différence $\mathcal{D}(n)$ entre taille moyenne d'un trie sous modèle de Poisson et d'un arbre des suffixes sous modèle de Bernoulli ont été bornées. Le comportement asymptotique de la taille moyenne d'un trie sous modèle sans mémoire est connu et nous avons :

Théorème (taille) *Le comportement asymptotique de l'espérance de la taille d'un arbre des suffixes construit sur les* n *premiers suffixes d'un mot produit par une source sans mémoire* (p, q) *avec* $p \geq q$ *et* $0.5 \leq p \leq 0.54$ *est donné quand* n *tend vers l'infini par :*

$$\mathbb{E}_n(S) = \frac{n}{h}(1 + \epsilon'(n)) + O(n^{0.85}),$$

où $h = -p\log p - q\log q$ *est l'entropie de la source et* ϵ' *une fonction oscillante de très faible amplitude (inférieure à* 10^{-5}*).*

Dans un premier temps, il nous faut étendre l'intervalle de validité du comportement asymptotique sur la probabilité p au-delà que ce maigre $[0.5, 0.54]$. Le principal problème est la contribution des motifs intermédiaires apériodiques. Deux approches sont possibles : soit affiner les ensembles sur lesquels les majorations sont effectuées, soit proposer des méthodes de majoration moins brutales que celle par le produit du nombre d'objets dans l'ensemble par le maximum de la quantité à sommer sur les objets de l'ensemble.

Dans le chapitre 5, une méthode plus fine est proposée pour déterminer le comportement asymptotique de la profondeur typique dans un arbre des suffixes construit à partir d'un texte engendré par une source markovienne d'ordre 1. Les moyennes de la profondeur typique D_n d'un arbre des suffixes et de la longueur de cheminement sont liées par la relation $\mathbb{E}(L_n) = n\mathbb{E}(D_n)$ et le chapitre 5 contient donc un résultat sur la longueur de cheminement. La méthode du chapitre 5 permet d'obtenir un résultat pour la longueur de cheminement pour toute probabilité p dans l'intervalle $[0, 1]$. La méthode apporte plus de précision sur la probabilité des motifs périodiques et apériodiques. En suivant la méthode du chapitre 6, l'intervalle de validité de la probabilité p pour lequel on connait le comportement asymptotique de la moyenne de la taille devrait lui aussi être étendu à l'intervalle $[0, 1]$.

La méthode présentée dans ce chapitre consiste en un contrôle du comportement asymptotique de la différence entre la moyenne du paramètre (taille, longueur de cheminement) pour un trie et pour un arbre des suffixes. Elle peut être appliquée à d'autres paramètres des arbres des suffixes mentionnés dans le chapitre 1. Le cas de la hauteur est particulier et d'autres techniques devront probablement être mises en œuvre pour obtenir des résultats significatifs. Il existe déjà plusieurs résultats sur la hauteur des arbres des suffixes [Szp91, AS92, DSR92].

Une autre étape consiste à passer à un modèle de source encore plus général que les ources markoviennes du chapitre 5 : les sources dynamiques introduites par Vallée [Val01]. Les outils changent et passent de l'analyse combinatoire et des séries génératrices à l'analyse fonctionnelle et aux valeurs propres des opérateurs.

Chapitre 3

Variances

Nous étudions la variance de la taille et de la longueur de cheminement d'un trie dans un modèle de source sans mémoire. Nous appliquons la transformée de Mellin de manière originale avec une méthode en « zigzag » pour montrer que la taille se comporte asymptotiquement de façon linéaire et la longueur de cheminement en $\Theta(n \log n)$ sous un modèle de Bernoulli. Dans une seconde partie nous comptons les textes avec certaines contraintes sur le nombre d'occurrences de deux motifs distincts w et w'.

We study the variance of size and path length in tries for a memoryless source model. We use a new method based on the Mellin transform called Mellin back-and-forth to prove that asymptotically the size behaves linearly and the external path length behaves as $\Theta(n \log n)$ in a Bernoulli model. In a second part, we count texts with some constraint on the number of occurrences of two patterns w and w'.

Sommaire

3.1 Introduction

Nous utilisons une nouvelle méthode pour trouver le comportement asymptotique de la longueur de cheminement d'un trie construit sur des chaînes produites par une source sans mémoire biaisée et pour un nombre de chaînes qui suit une loi de Poisson de paramètre z puis sous un modèle de Bernoulli. Nous établissons ensuite que pour un trie construit sur exactement n chaînes, le comportement asymptotique de la longueur de cheminement est en $\Theta(n \log n)$. Cette même méthode permet de retrouver le comportement asymptotique linéaire de la variance de la taille d'un trie construit sur un nombre de chaînes suivant une loi de Poisson de paramètre z et sur des chaînes engendrées par une source sans mémoire biaisée.

Dans une seconde partie, nous utilisons les méthodes de décomposition combinatoire introduites par Guibas et Odlyzko et qui nous ont servi dans le chapitre précédent pour trouver l'expression de séries génératrices qui comptent les co-occurrences de deux motifs w et w' dans les textes. Les séries génératrices ainsi obtenues sont des fractions rationnelles qui s'écrivent en fonction des polynômes d'auto-corrélation $c[w](z)$ de w et $c[w'](z)$ de w' et des polynômes de corrélation croisés $c[w, w'](z)$ et $c[w', w](z)$ de w et w'.

Les variances de la taille et de la longueur de cheminement d'un trie ont été relativement peu étudiées. Pourtant la connaissance de la variance aide à cerner si un paramètre est proche de sa moyenne ou non. Par exemple la variance de la longueur de cheminement d'un arbre permet de savoir si cette classe d'arbres est bien équilibrée. La variance de la taille d'un trie a été déterminée par Jacquet et Régnier [JR87, JR88, RJ89] dans le cas sans mémoire. Kirschenhofer et Prodinger [KP91] se sont servis de formules dues à Ramanujan[17] pour annuler certains termes et trouver la variance de la taille d'un trie dans un modèle symétrique. Kirschenhofer, Prodinger et Szpankowski [KPS89] ont obtenu le comportement asymptotique de la variance de la longueur de cheminement dans un trie pour un modèle de source symétrique. Si la moyenne de la profondeur typique D_n et de la longueur de cheminement L_n sont reliées par la formule

$$n\mathbb{E}(D_n) = \mathbb{E}(L_n),$$

il n'en est rien pour les variances de ces deux paramètres. Donc les résultats plus nombreux sur la variance de la profondeur typique ne peuvent servir à récupérer la variance de la longueur de cheminement. De plus, pour la variance, les comportements asymptotiques dans le cas d'une source symétrique et dans le cas d'une source biaisée sont assez différents. Sur l'étude de la taille, notre approche donne pour l'instant un résultat plus faibles que ceux de Jacquet et Régnier car nous ne sommes pas à même de donner explicitement le coefficient du terme linéaire. Notre méthode présente cependant plusieurs avantages : contrairement à Kirschenhofer et Prodinger, nous obtenons des résultat dans le cas biaisé et pas uniquement dans le cas symétrique. Nous n'avons pas besoin, comme Jacquet et Régnier, de déterminer la distribution limite de la taille pour trouver sa variance. Notre approche présente donc un calcul direct de la variance de la taille et de la longueur de cheminement dans un trie dans le cas biaisé. De plus nos résultats sur le comportement asymptotique de la longueur de cheminement d'un trie (sous un modèle de Poisson et de Bernoulli) semblent être les premiers à fournir une preuve complète du terme de deuxième ordre. Les résultats sur la longueur de cheminement sont résumés par les deux théorèmes suivants où la variance est notée $\mathbb{V}$ et le modèle sur le nombre de textes est spécifié en indice ($\mathcal{P}(z)$ pour le modèle de Poisson et n pour le modèle de Bernoulli) :

Théorème 6 *Le comportement asymptotique de la variance de la longueur de cheminement d'un trie construit sur des textes engendrés par un modèle de source sans mémoire (p, q) est, pour un modèle de Poisson de paramètre z*

$$\mathbb{V}_{\mathcal{P}(z)}(L) = \frac{1}{h^2} z \log^2 z + K_5 z \log z + O(z), \quad \textit{et pour un modèle de Bernoulli,} \tag{3.1}$$

$$\mathbb{V}_n(L) = K_6 n \log n + O(n), \quad \textit{où} \tag{3.2}$$

$$K_5 := \frac{2h_2}{h^3} + \frac{2(1+\gamma)}{h^2} - \frac{1}{h}, \quad K_6 := \frac{h_2 - h^2}{h^3} \quad \textit{et} \quad h_2 := p\log^2 p + q \log^2 q. \tag{3.3}$$

La variance de la taille et de la longueur de cheminement dans un arbre des suffixes n'a fait l'objet, à notre connaissance, d'aucune étude. Ce chapitre apporte les premières pierres à cette étude. Le plan de travail futur sera détaillé dans la conclusion.

Dans le chapitre 1, nous exprimions la moyenne de la taille et de la longueur de cheminement comme la somme sur tous les motifs de certaines quantités liées au paramètre N_w. Dans les sections 3.2 et 3.3 de ce chapitre, nous appliquons les méthodes de décomposition du chapitre 1

[17] Srinivasa Aiyangar Ramanujan, mathématicien indien (1887–1920)

pour exprimer la variance de la taille et de la longueur de cheminement d'un trie comme la somme sur toutes les paires de motifs de certaines quantités liées au paramètre N_w. Le comportement asymptotique de chacune des sommes de la décomposition combinatoire de la variance de la taille et de la longueur de cheminement est déterminé à l'aide de la transformation de Mellin. Une utilisation répétitive des théorèmes de Mellin « direct » et « indirect » est présentée à la page 71. Elle permet de déterminer le comportement asymptotique de sommes compliquées.

Ensuite nous obtenons, par une méthode similaire à celle de Guibas et Odlyzko [GO81b], les quatre séries génératrices qui comptent le nombre de textes de taille donnée avec 0 ou 1 occurrence d'un motif w et d'un autre motif w'. La formulation de ces séries génératrices et plus particulièrement de leur dénominateur rend l'extraction des singularités, premier pas vers la détermination du comportement asymptotique de la variance de la taille et de la longueur de cheminement dans un arbre des suffixes, difficile.

3.2 Variance de la taille des tries

Dans cette section nous montrons que le comportement asymptotique de la variance de la taille d'un trie dans un modèle de source sans mémoire pour la génération des textes et sous un modèle de Poisson pour le nombre de textes est linéaire.

La taille S d'un trie a déjà été définie à la page 20 du chapitre 1 comme le nombre de nœuds internes du trie. On note X l'ensemble des chaînes à partir duquel le trie est construit. Il y a un nœud interne en une position donnée dans le trie si le motif qui est lié à ce nœud apparaît à au moins deux reprises comme préfixe des mots de X. La variance est notée $\mathbb{V}$. La variance est définie par :

$$\mathbb{V}(S) := \mathbb{E}(S^2) - \mathbb{E}^2(S).$$

Le comportement asymptotique du carré de l'espérance est le carré du comportement asymptotique de l'espérance, or nous connaissons un développement asymptotique complet de l'espérance de la taille d'un trie. Nous savons ainsi obtenir le comportement asymptotique du carré de la taille. Le point le plus délicat est de trouver le comportement asymptotique de la moyenne de S^2. Nous décomposons combinatoirement les paires de motifs en deux ensembles selon que ces motifs sont préfixes l'un de l'autre ou non. Cela se traduit au niveau du trie en relation d'ancestralité des nœuds associés aux motifs. Nous obtenons

Théorème 7 *Le comportement asymptotique de la variance de la taille dans un trie sous modèle de Poisson de paramètre z et sous un modèle de source sans mémoire vaut*

$$\mathbb{V}_{\mathcal{P}(z)}(S^2) = O(z).$$

Dans les premières sections, nous trouvons des formulations explicites des différentes contributions à l'espérance du carré de la taille sous le modèle de Poisson de paramètre z. La technique de la transformation de Mellin nous sert ensuite à obtenir le comportement asymptotique de chacune des sommes qui entrent en jeu. Nous introduisons une nouvelle méthode issue de la théorie de Mellin pour recouvrer le comportement asymptotique de certaines sommes. Cette méthode de « Mellin en zig-zag » est détaillée à la page 71 et appliquée dans la même section.

Le carré de la taille s'écrit

$$\begin{aligned} S^2 &= \sum_{w\in\mathcal{A}^\star} [\![N_w \geq 2]\!]^2 + \sum_{w,w'\in\mathcal{A}^\star, w\neq w'} [\![N_w \geq 2]\!][\![N_{w'} \geq 2]\!] \\ &= \sum_{w\in\mathcal{A}^\star} [\![N_w \geq 2]\!] + \sum_{w,w'\in\mathcal{A}^\star, w\neq w'} (1 - [\![N_w = 0]\!] - [\![N_w = 1]\!]).(1 - [\![N_{w'} = 0]\!] - [\![N_{w'} = 1]\!]). \end{aligned}$$

La probabilité que le motif w apparaisse en préfixe d'au moins deux textes de l'ensemble de base X sous un modèle de Poisson de paramètre z vaut $1 - (1 + zp_w)\exp(-zp_w)$. En prenant l'espérance du carré de la taille, on obtient :

$$\begin{aligned} \mathbb{E}_{\mathcal{P}(z)}(S^2) = &\sum_{w\in\mathcal{A}^\star} (1 - (1 + zp_w)\exp(-zp_w)) \\ &+ \sum_{(w,w')\in\mathcal{A}^\star\times\mathcal{A}^\star, w\neq w'} 1 - (1 + zp_w)\exp(-zp_w) - (1 + zp_{w'})\exp(-zp_{w'}) \\ &+ \mathbb{E}_{\mathcal{P}(z)}\left([\![N_w = 0]\!][\![N_{w'} = 0]\!]\right) + \mathbb{E}_{\mathcal{P}(z)}\left([\![N_w = 1]\!][\![N_{w'} = 1]\!]\right) \\ &+ \mathbb{E}_{\mathcal{P}(z)}\left([\![N_w = 0]\!][\![N_{w'} = 1]\!]\right) + \mathbb{E}_{\mathcal{P}(z)}\left([\![N_w = 1]\!][\![N_{w'} = 0]\!]\right). \end{aligned} \tag{3.4}$$

Dans la suite du chapitre, on note

$$\Theta_1(z) := \sum_{w\in\mathcal{A}^\star} (1 - (1 + zp_w)\exp(-zp_w)), \tag{3.5}$$

cette somme sera appelée « simple. » Il nous faut maintenant calculer les espérances des quatre indicatrices qui entrent en jeu dans la seconde somme (somme sur tous les couples de motifs distincts), c'est-à-dire

$$\mathbb{E}_{\mathcal{P}(z)}([\![N_w = a]\!][\![N_{w'} = b]\!]) = \mathbb{P}_{\mathcal{P}(z)}(N_w = a \cap N_{w'} = b), \tag{3.6}$$

avec a et b appartenant à $\{0, 1\}$. Si, par exemple, w est préfixe de w', les deux événements $[\![N_w = 0]\!]$ et $[\![N_{w'} = 0]\!]$ sont fortement liés : si w n'apparaît pas comme préfixe d'un texte de S alors w' n'apparaitra jamais non plus.

L'étude de l'espérance se scinde naturellement en deux selon que l'un des deux motifs soit préfixe de l'autre ou non. Si aucun des deux mots w et w' n'est préfixe l'un de l'autre, on dit alors qu'ils sont *cousins*. Nous introduisons la notation $w \wedge w'$ pour désigner le plus long préfixe commun de w et w'. Il y a quatre espérances à calculer pour les deux ensembles de motifs : cousins ou préfixes ce qui fait 8 cas à étudier. Pour simplifier l'écriture, nous gardons en mémoire que le modèle sur le nombre de chaînes est ici de Poisson et qu'il n'est plus nécessaire le préciser en indice des différentes espérances et probabilités.

3.2.1 Expression de la contribution des cousins

Si $u := w \wedge w'$ est différent de w et w' alors les deux chaînes w et w' sont cousines. La figure 3.1 illustre la position des nœuds associés aux mots u, w et w' dans un trie par rapport à la racine (en noir). Les quatre probabilités de (3.6) qui nous intéressent sont exprimées à l'aide de probabilités conditionnelles. Les résultats qui sont obtenus sont présentés dans le cas particulier d'un modèle sans mémoire, mais la preuve s'étend à un modèle plus général.

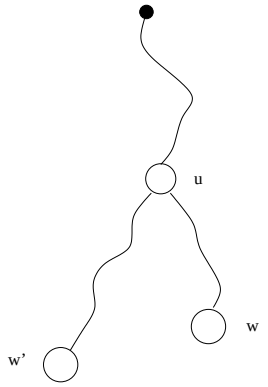


FIG. 3.1 – w et w' sont cousins

➜ $N_w = 0$ **et** $N_{w'} = 0$

On traite en détail le cas de la probabilité $\mathbb{P}(N_w = 0 \cap N_{w'} = 0)$. On a

$$\mathbb{P}(N_w = 0 \cap N_{w'} = 0) = \sum_{n \geq 0} \mathbb{P}(N_w = 0 \cap N_{w'} = 0 \,|\, N_u = n)\mathbb{P}(N_u = n).$$

Puisqu'on s'est placé sous un modèle de Poisson pour le nombre de chaînes dans l'ensemble de base X, la variable aléatoire N_u suit une loi de Poisson de paramètre zp_u. La probabilité conditionnelle se traduit comme la probabilité que parmi les n chaînes qui commencent par le préfixe u, aucune ne commence ni par le préfixe w, ni par le préfixe w'. Le nombre de chaînes commençant par u est fixé, on a donc

$$\mathbb{P}(N_w = 0 \cap N_{w'} = 0 \,|\, N_u = n) = (1 - p_{w|u} - p_{w'|u})^n,$$

où $p_{w|u}$ est la probabilité qu'un motif commence par le préfixe w sachant qu'il commence déjà par le motif u. Ainsi sous le modèle de Poisson de paramètre z, la probabilité que ni le motif w, ni le motif w' n'apparaissent en préfixe des textes de l'ensemble de base :

$$\begin{aligned}
\mathbb{P}(N_w = 0 \cap N_{w'} = 0) &= \sum_{n \geq 0} \mathbb{P}(N_w = 0 \cap N_{w'} = 0 \,|\, N_u = n)\mathbb{P}(N_u = n) \\
&= \sum_{n \geq 0} (1 - p_{w|u} - p_{w'|u})^n \frac{(p_u z)^n \exp(-zp_u)}{n!} \\
&= \exp(-zp_u p_{w|u}) \exp(-zp_u p_{w'|u}).
\end{aligned}$$

Dans le cas où les textes sont engendrés par une source sans mémoire, la probabilité conditionnelle $p_{w|u}$ vaut p_w/p_u et on a

$$\mathbb{P}(N_w = 0 \cap N_{w'} = 0) = \exp(-zp_w) \exp(-zp_{w'}). \tag{3.7}$$

➜ $N_w = 1$ **et** $N_{w'} = 0$

On réutilise la même idée que dans le paragraphe précédent, à savoir déterminer la probabilité conditionnelle d'avoir une unique chaîne qui commence par w, de n'avoir aucune chaîne qui commence par w' en sachant que n mots commencent par le préfixe u. Cette condition se traduit

par : parmi les n chaînes qui commencent par u, une seule commence par w et les $n-1$ autres ne commencent ni par w, ni par w'. Soit

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 0 \,|\, N_u = n) = np_{w|u}(1 - p_{w|u} - p_{w'|u})^{n-1},$$

pour n strictement positif.

$$\begin{aligned}\mathbb{P}(N_w = 1 \cap N_{w'} = 0) &= \sum_{n\geq 1} \mathbb{P}(N_w = 1 \cap N_{w'} = 0 \,|\, N_u = n)\mathbb{P}(N_u = n) \\ &= \sum_{n\geq 1} np_{w|u}(1 - p_{w|u} - p_{w'|u})^{n-1} \frac{(zp_u)^n \exp(-zp_u)}{n!} \\ &= zp_u p_{w|u} \exp(-zp_u p_{w|u}) \exp(-zp_u p_{w'|u}).\end{aligned}$$

Pour les textes engendrés par une source sans mémoire, on a

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 0) = zp_w \exp(-zp_w) \exp(-zp_{w'}), \tag{3.8}$$

et de manière symétrique,

$$\mathbb{P}(N_w = 0 \cap N_{w'} = 1) = zp_{w'} \exp(-zp_w) \exp(-zp_{w'}). \tag{3.9}$$

➜ $N_w = 1$ **et** $N_{w'} = 1$

On cherche maintenant la probabilité que parmi les n chaînes qui commencent par u, il y en ait exactement une qui commence par w et exactement une qui commence par w'. Cela se traduit par

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 1 \,|\, N_u = n) = \binom{n}{1} p_{w|u} \binom{n-1}{1} p_{w'|u}(1 - p_{w|u} - p_{w'|u})^{n-2}.$$

On obtient alors facilement

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 1) = (zp_u)^2 p_{w|u} p_{w'|u} \exp(-zp_u p_{w|u}) \exp(-zp_u p_{w'|u}).$$

Et ce résultat s'exprime pour des sources sans mémoire par :

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 1) = (zp_w \exp(-zp_w))(zp_{w'} \exp(-zp_{w'})). \tag{3.10}$$

En assemblant les résultats que l'on vient de trouver, il vient

$$\begin{aligned}\Theta_2(z) := &\sum_{(w,w') \text{ cousins}} 1 - (1 + zp_w)\exp(-zp_w) - (1 + zp_{w'})\exp(-zp_{w'}) + \mathbb{E}\left([\![N_w = 0]\!][\![N_{w'} = 0]\!]\right) \\ &+ \mathbb{E}\left([\![N_w = 1]\!][\![N_{w'} = 1]\!]\right) + \mathbb{E}\left([\![N_w = 0]\!][\![N_{w'} = 1]\!]\right) + \mathbb{E}\left([\![N_w = 1]\!][\![N_{w'} = 0]\!]\right) \\ = &\sum_{(w,w') \text{ cousins}} (1 - (1 + zp_w)\exp(-zp_w))(1 - (1 + zp_{w'})\exp(-zp_{w'})).\end{aligned}$$

Si w et w' sont cousins, il existe un motif $u = w \wedge w'$ qui est le préfixe commun aux deux mots. Les motifs w et w' s'écrivent sous la forme $w = u.0.v$ et $w' = u.1.v'$ (ou l'inverse, le plus important étant que la lettre après u ne soit pas la même pour les deux mots). Ainsi en tenant compte de la symétrie entre les deux motifs, on obtient

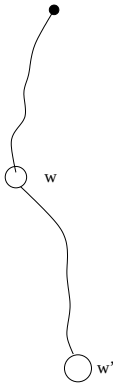


FIG. 3.2 – w est un préfixe de w'

Lemme 17 *La contribution des motifs cousins à l'espérance du carré de la taille d'un trie s'écrit dans un modèle sans mémoire de génération des textes et dans un modèle de Poisson de paramètre z sur le nombre de textes*

$$\Theta_2(z) = 2 \sum_{(u,v,v')\in(\mathcal{A}^\star)^3} (1-(1+zp_{u0v})\exp(-zp_{u0v}))(1-(1+zp_{u1v})\exp(-zp_{u1v})).$$

3.2.2 Expression de la contribution des préfixes

Dans cette section, nous traitons du cas où l'une des deux chaînes est préfixe de l'autre. On prend la convention que w est un préfixe de w'. Cela se traduit en utilisant notre notation par $w \wedge w' = w$, la notation standard $w \sqsubset w'$ sert à exprimer que w est préfixe de w'. La figure 3.2 éclaircit cette configuration dans le trie formé sur un ensemble des mots X dont au moins une chaîne commence par w et au moins une chaîne par w'. Nous tirons avantage des incompatibilités entre deux événements pour exprimer les probabilités recherchées.

➜ $N_w = 0$ **et** $N_{w'} = 0$

Nous calculons la probabilité qu'aucune chaîne ne commence par w, ni par w'. Le motif w est un préfixe de w', et donc si aucune chaîne de l'ensemble X ne commence par w, nécessairement, aucune chaîne ne peut commencer par w'. Il nous suffit de déterminer la probabilité qu'aucune chaîne ne commence par w. Dans un modèle de Poisson, on obtient

$$\mathbb{P}(N_w = 0 \cap N_{w'} = 0) = \mathbb{P}(N_w = 0) = \exp(-zp_w).$$

➜ $N_w = 1$ **et** $N_{w'} = 0$

La probabilité qu'une seule chaîne de l'ensemble de base commence par w et aucune par w' est la probabilité que la seule chaîne de X qui commence par w ne commence pas par w'. On a recours à une expression conditionnelle :

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 0) = \mathbb{P}(N_{w'} = 0 | N_w = 1)\mathbb{P}(N_w = 1),$$

or la probabilité qu'un mot qui a w comme préfixe n'ait pas aussi w' comme préfixe est $1-p_{w|w'}$. D'où

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 0) = (1 - p_{w'|w}) z p_w \exp(-z p_w).$$

Pour une source sans mémoire, on a $p_{w'|w} = p_{w'}/p_w$, ainsi :

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 0) = z(p_w - p_{w'}) \exp(-z p_w).$$

➜ $N_w = 0$ **et** $N_{w'} = 1$

Ce cas est evacué rapidement puisque si aucun mot de l'ensemble ne commence par w, il ne peut y en avoir qui commence par w' (w est un préfixe de w'). Ainsi

$$\mathbb{P}(N_w = 0 \cap N_{w'} = 1) = 0.$$

➜ $N_w = 1$ **et** $N_{w'} = 1$

On exprime la probabilité que parmi les mots de l'ensemble X, un seul commence par w et un seul par w'. Mais la relation entre w et w' nous dit que la condition se traduit par : l'unique chaîne qui commence par w doit aussi commencer par w'. Cette condition se récrit sous la forme

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 1) = \mathbb{P}(N_{w'} = 1 \,|\, N_w = 1)\mathbb{P}(N_w = 1)$$

On applique ensuite l'expression de la probabilité sous le modèle de Poisson et pour obtenir

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 1) = z p_w p_{w'|w} \exp(-z p_w),$$

ainsi, dans le cas d'une source sans mémoire, la probabilité s'exprime

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 1) = z p_{w'} \exp(-z p_w).$$

La somme sur les motifs préfixes se décompose en deux, selon que w est préfixe de w' ou w' est préfixe de w. Les deux cas sont symétriques et en remplacant avec les résultats de cette section, on a

$$\begin{aligned}\Theta_3(z) :=& \sum_{(w,w') \text{ préfixes}} 1 - (1 + z p_w)\exp(-z p_w) - (1 + z p_{w'})\exp(-z p_{w'}) + \mathbb{E}\left([\![N_w = 0]\!][\![N_{w'} = 0]\!]\right) \\ &+ \mathbb{E}\left([\![N_w = 1]\!][\![N_{w'} = 1]\!]\right) + \mathbb{E}\left([\![N_w = 0]\!][\![N_{w'} = 1]\!]\right) + \mathbb{E}\left([\![N_w = 1]\!][\![N_{w'} = 0]\!]\right) \\ =& 2 \sum_{w \sqsubset w'} (1 - (1 + z p_{w'})\exp(-z p_{w'})),\end{aligned}$$

où $\{w \sqsubset w'\}$ est l'ensemble des couples de motifs (w, w') pour lesquels w est un préfixe de w'.

La somme $\Theta_3(z)$ se simplifie puisqu'à w' fixé (et pour w préfixe de w') il ne peut y avoir que $|w'|$ choix pour le motif w (les ascendants du nœud représentant w'). Et puisque la quantité à sommer ne dépend pas de w, on obtient

Lemme 18 *La contribution des motifs préfixes à l'espérance du carré de la taille du trie sous modèle de Poisson de paramètre z s'écrit*

$$\Theta_3(z) = 2 \sum_{w \in \mathcal{A}^\star} |w|(1 - (1 + z p_w)\exp(-z p_w)).$$

La notation $Q_w(z) := 1 - (1 + zp_w)\exp(-zp_w)$ va permettre de raccourcir les expressions qui entrent en jeu et de formuler l'espérance du carré de la taille de façon plus agréable

$$\begin{aligned}\mathbb{E}(S^2) &= \Theta_1(z) + \Theta_2(z) + \Theta_3(z)\\ &= \sum_{w\in\mathcal{A}^\star} Q_w(z) + \sum_{(u,v,v')\in(\mathcal{A}^\star)^3} Q_{u0v}(z)Q_{u1v'}(z) + 2\sum_{w\in\mathcal{A}^\star} |w|Q_w(z),\end{aligned} \tag{3.11}$$

Les sections suivantes servent à déterminer le comportement asymptotique de $\mathbb{E}(S^2)$. On cherche le comportement asymptotique des trois sommes $\Theta_1(z)$, $\Theta_2(z)$ et $\Theta_3(z)$ qui entrent en jeu dans l'équation (3.11). Le comportement asymptotique de la somme $\Theta_1(z)$ quand z tend vers l'infini est déjà connu puisqu'il s'agit de la moyenne de la taille du trie. On regarde la somme $\Theta_2(z)$ sur les motifs cousins avant de s'occuper de la somme $\Theta_3(z)$.

3.2.3 Contribution asymptotique des cousins

Dans cette section, nous déterminons le comportement asymptotique de la contribution des motifs cousins à l'espérance du carré de la taille d'un trie, c'est-à-dire le comportement asymptotique de la somme $\Theta_2(z)$ quand z tend vers l'infini. Nous introduisons une variation sur le thème de la transformation de Mellin avec la transformation en « zigzag. » D'habitude, pour obtenir le comportement asymptotique d'une somme, on détermine dans un premier temps la transformée de Mellin de cette somme. Puis on extrait de l'information sur les pôles, leur localisation et la valeur du résidu en ces pôles. Enfin, le théorème de Mellin « inverse » offre le comportement asymptotique de la somme. L'application en zigzag va répéter cette méthode pour des sommes dépendant de plusieurs paramètres. Voici un résumé du fonctionnement de la transformation de Mellin en « zig-zag » :

Application en *zigzag*

Regardons le principe de l'application en *zigzag* de la théorie de Mellin pour une fonction $F(z)$ et une somme

$$\sum_{u\in\mathcal{A}^\star}\sum_{(v,v')\in(\mathcal{A}^\star)^2} F(zp_up_v)F(zp_up_{v'}).$$

Notre objectif est de trouver le comportement asymptotique de cette somme quand z tend vers l'infini. Voici les étapes :

- trouver le comportement asymptotique de la somme $\sum_{v\in\mathcal{A}^*} F(zp_v)$ par la méthode « classique » ;
- puis le comportement asymptotique du produit $G(z) := \sum_{v,v'\in\mathcal{A}^*} F(zp_v)F(zp_{v'})$ (c'est le produit des deux comportements asymptotiques) ;
- à partir de l'asymptotique de $G(z)$, on connait le développement singulier de $G^\star(s)$ (application du théorème de Mellin « direct ») ;
- on détermine le développement singulier de la transformée de Mellin de $\sum_u G(zp_u)$;
- et conclut par une application du théorème de Mellin « inverse » pour obtenir le comportement asymptotique recherché.

Nous allons maintenant appliquer cette méthode générique à la fonction

$$f := z \mapsto 1 - (1 + z)\exp(-z).$$

La somme à étudier est

$$\sum_{(u,v,v')\in(\mathcal{A}^\star)^3} Q_{u0v}(z)Q_{u1v'}(z).$$

La première étape de la méthode en « zigzag » est la détermination du comportement asymptotique de

$$\Upsilon_{\mathtt{0}}(z) := \sum_{v\in\mathcal{A}^*} Q_{\mathtt{0}.v}(z) = \sum_{v\in\mathcal{A}^\star} f(zp_{\mathtt{0}.v}).$$

La source qui engendre les textes est sans mémoire et ainsi $p_{\mathtt{0}.v} = pp_v$. La bande fondamentale de la fonction f est $\langle -2, 0\rangle$, la bande de convergence de la série de Dirichlet est $\langle -\infty, -1\rangle$. Par conséquent la transformée de Mellin est définie dans la bande $\langle -2, -1\rangle$ et y vaut :

$$f^\star(s) = -(s+1)\Gamma(s) \text{ d'où } \Upsilon_{\mathtt{0}}^\star(s) = -(s+1)\Gamma(s)\sum_{v\in\mathcal{A}^\star} p^{-s}p_v^{-s} = -\frac{(s+1)p^{-s}\Gamma(s)}{1-(p^{-s}+q^{-s})}.$$

Il y a quatre types de pôles de $\Upsilon_{\mathtt{0}}^\star(s)$ qui entrent en compte dans la détermination du comportement asymptotique au voisinage de $\Upsilon_{\mathtt{0}}(z)$ jusqu'au terme constant :

- pôle simple en $s = -1$,
- si la source est périodique (c'est-à-dire si $\log p/\log q \in \mathbb{Q}$), des pôles simples en les $-1+i\alpha_k$ avec $\alpha_k = 2k\pi/(\log p - \log q)$ et $k \in \mathbb{Z}^\star$,
- pôles en les solutions de $p^{-s} + q^{-s} = 1$ dans la bande $\langle -1, 0\rangle$;
- pôle en 0.

L'ensemble des racines de l'équation $p^{-s} + q^{-s} = 1$ dans la bande $\langle -1, 0\rangle$ est défini comme l'ensemble des β_j pour $j \in \mathcal{J}$. On sait d'après [FFH86] que les pôles de l'équation $p^{-s}+q^{-s} = 1$ sont tous simples et de plus la fonction Gamma ne peut s'annuler en ces points. Je fournis de plus le lemme suivant

Lemme 19 *Si la source est apériodique, il n'existe pas de solution de partie réelle nulle à l'équation* $p^{-s} + q^{-s} = 1$.

Preuve : Si la partie réelle de s est nulle, p^{-s} et q^{-s} ont tous deux module 1, ils se situent donc sur le cercle trigonométrique. La somme de leurs parties imaginaires s'annulent, les points qui représentent ces deux nombres complexes sont donc symétriques, soit par rapport à l'origine, soit par rapport à l'axe réel. Dans le premier cas, leurs parties réelles se compensent aussi et donc on ne peut avoir la condition $p^{-s} + q^{-s} = 1$; dans le second cas leurs parties réelles sont égales et il faut donc avoir $\Re(p^{-s}) = 1/2$. En choisissant p^{-s} comme celui de deux avec partie imaginaire positive, on a $-\Im(s)\log p = \pi/3 + 2k\pi$ et $-\Im(s)\log q = -\pi/3 + 2k'\pi$ avec $k, k' \in \mathbb{Z}$. Mais alors le rapport des deux logarithmes est rationnel, ce qui contredit notre hypothèse d'apériodicité. ◀

Ce lemme nous garantit que, dans le cas d'une source apériodique, le seul pôle de la transformée de Mellin de partie réelle nulle est le pôle simple en zéro. Après une application du théorème de Mellin « inverse », le comportement asymptotique de $\Upsilon_{\mathtt{0}}(s)$ dans le cas d'une source périodique s'exprime par

$$\begin{aligned}\frac{zp}{h}\left(1 - \sum_{k\in\mathbb{Z}^\star} i\alpha_k\Gamma(-1+i\alpha_k)z^{-i\alpha_k}\right) - \sum_{j\in\mathcal{J}}\frac{(\beta_j+1)p^{-\beta_j}\Gamma(\beta_j)}{H_{\beta_j}}z^{-\beta_j}\\ - \sum_{i\in\mathcal{I}}\frac{(\beta_i+1)p^{-\beta_i}\Gamma(\beta_i)}{H_{\beta_i}}z^{-\beta_i} - 1 + o(1),\end{aligned} \tag{3.12}$$

où $H_{\beta_j} := -p^{-\beta_j}\log p - q^{-\beta_j}\log q$ et β_i pour $i \in \mathcal{I}$ est l'ensemble des solutions de l'équation $p^{-s} + q^{-s} = 1$ situées sur l'axe $\Re = 0$. Pour une source apériodique, on a une expression

asymptotique plus simple :

$$\frac{zp}{h} - \sum_{j\in\mathcal{J}} \frac{(\beta_j+1)p^{-\beta_j}\Gamma(\beta_j)}{H_{\beta_j}} z^{-\beta_j} - 1 + o(1).$$

L'auteur espère que le lecteur lui pardonnera de ne traîter que le cas d'une source apériodique. Le cas périodique ne fait que rendre les calculs un peu plus compliqués mais ne rajoute aucune difficulté conceptuelle, ni n'exhibe de comportement original. Le comportement asymptotique de la fonction $\Upsilon(z) := \Upsilon_0(z)\Upsilon_1(z)$ est, on l'a déjà dit, le produit des deux comportements asymptotiques, c'est-à-dire dans le cas apériodique

$$\begin{aligned}\frac{z^2pq}{h^2} &- \sum_{j\in\mathcal{J}} \frac{(\beta_j+1)\Gamma(\beta_j)(qp^{-\beta_j}+pq^{-\beta_j})}{hH_{\beta_j}} z^{-\beta_j+1} \\ &+ \sum_{(j,k)\in\mathcal{K}} \frac{(\beta_j+1)(\beta_k+1)p^{-\beta_j}q^{-\beta_k}\Gamma(\beta_j)\Gamma(\beta_k)}{H_{\beta_k}H_{\beta_j}} z^{-\beta_j-\beta_k} - \frac{z}{h} + o(z),\end{aligned}$$

où $\mathcal{K}$ est l'ensemble des paires d'éléments de $\mathcal{J}$ qui vérifient $\Re(\beta_j+\beta_k) \leq -1$. Nous avons tronqué les termes sous-linéaires du développement asymptotique de $\Upsilon(z)$.

On suit la méthode de Mellin « zigzag » donnée précédemment. Il n'y a pas de pôles double dans le développement singulier de la transformée de Mellin $\Upsilon^\star(s)$ (car il n'y a pas de termes faisant intervenir la fonction log dans le développement asymptotique) et la borne droite de la bande fondamentale est -2. Le développement singulier de $\Upsilon^\star(s)$ en les singularités entre -2 et -1 vaut

$$\begin{aligned}&- \frac{pq}{(s+2)h^2} + \sum_{j\in\mathcal{J}} \frac{(\beta_j+1)\Gamma(\beta_j)(qp^{-\beta_j}+pq^{-\beta_j})}{hH_{\beta_j}} \frac{1}{s+1-\beta_j} \\ &- \sum_{(j,k)\in\mathcal{K}} \frac{(\beta_j+1)(\beta_k+1)p^{-\beta_j}q^{-\beta_k}\Gamma(\beta_j)\Gamma(\beta_k)}{H_{\beta_k}H_{\beta_j}} \frac{1}{s-\beta_j-\beta_k} + \frac{1}{h(s+1)}.\end{aligned}$$

La transformée de Mellin de la somme $(1/2)\Theta_2(z)$ vaut $\sum_{u\in\mathcal{A}^\star} p_u^{-s}\Upsilon^\star(s)$. Pour obtenir le développement singulier de $\Theta_2^\star(s)$ en les singularités entre -2 et -1, il faut rajouter le pôle en -1 de la somme de Dirichlet, qui crée un pôle double pour $\Theta_2^\star(s)$ en -1. Il n'y a aucun autre pôle pour la série de Dirichlet sur l'axe $\Re = -1$ en dehors du pôle en -1 (on a une source apériodique). Les autres solutions de $p^{-s}+q^{-s}=1$ sont de partie réelle plus grande que -1. Ainsi le développement singulier s'écrit

$$\begin{aligned}&- \frac{pq}{(s+2)h^2(1-(p^2+q^2))} + \sum_{j\in\mathcal{J}} \frac{(\beta_j+1)\Gamma(\beta_j)(qp^{-\beta_j}+pq^{-\beta_j})}{(1-(p^{1-\beta_j}+q^{1-\beta_j}))hH_{\beta_j}} \frac{1}{s+1-\beta_j} \\ &- \sum_{(j,k)\in\mathcal{K}} \frac{(\beta_j+1)(\beta_k+1)p^{-\beta_j}q^{-\beta_k}\Gamma(\beta_j)\Gamma(\beta_k)}{(1-(p^{-\beta_j-\beta_k}+q^{-\beta_j-\beta_k}))H_{\beta_k}H_{\beta_j}} \frac{1}{s-\beta_j-\beta_k} - \frac{1}{h^2(s+1)^2},\end{aligned} \tag{3.13}$$

et le comportement asymptotique de la somme $\Theta_2(z)$ au voisinage de l'infini résulte d'une application du théorème de Mellin « inverse. »

Lemme 20 *Pour une source apériodique et sous un modèle de Poisson de paramètre z, le comportement asymptotique de la somme $\Theta_2(z)$ quand z tend vers l'infini est*

$$\frac{2pqz^2}{h^2(1-(p^2+q^2))} - 2\sum_{j\in\mathcal{J}} \frac{(\beta_j+1)\Gamma(\beta_j)(qp^{-\beta_j}+pq^{-\beta_j})}{(1-(p^{1-\beta_j}+q^{1-\beta_j}))hH_{\beta_j}} z^{1-\beta_j} + 2\sum_{(j,k)\in\mathcal{K}} \frac{(\beta_j+1)(\beta_k+1)p^{-\beta_j}q^{-\beta_k}\Gamma(\beta_j)\Gamma(\beta_k)}{(1-(p^{-\beta_j-\beta_k}+q^{-\beta_j-\beta_k}))H_{\beta_k}H_{\beta_j}} z^{-\beta_j-\beta_k} - 2\frac{z\log z}{h^2} + O(z). \tag{3.14}$$

3.2.4 Contribution asymptotique des préfixes

Nous obtenons le comportement asymptotique de $\Theta_3(z)$, c'est-à-dire la contribution des motifs w et w' quand l'un est préfixe de l'autre. La somme à étudier est

$$\Theta_3(z) = 2\sum_{w\in\mathcal{A}^\star} |w| f(zp_w).$$

Les résultats du paragraphe précédent nous donnent, *via* la transformation de Mellin, le comportement asymptotique de la fonction $f := z \mapsto 1-(1+z)\exp(-z)$ quand z est au voisinage de l'infini. Sa bande fondamentale est $\langle -2, 0\rangle$ et sa transformée de Mellin vaut :

$$f^\star(s) = -(s+1)\Gamma(s).$$

Ainsi la transformée de Mellin de $\Theta_3(z)$ est

$$\Theta_3^\star(s) := -2(s+1)\Gamma(s)\sum_{n\geq 0} n \sum_{w\in\mathcal{A}^n} p_w^{-s}.$$

Dans le cas d'une source sans mémoire biaisée, la somme interne de la transformée de Mellin s'écrit

$$\sum_{n\geq 0} n \sum_{w\in\mathcal{A}^n} p_w^{-s} = \sum_{n\geq 0} n \sum_{i=0}^{n} \binom{n}{i} (p^i q^{k-i})^{-s} = \sum_{n\geq 0} n(p^{-s}+q^{-s})^n.$$

La bande de convergence de la transformée de Mellin est $\langle -\infty, -1\rangle$ pour que $p^{-s}+q^{-s} < 1$. Ainsi $\Theta_3^\star(s)$ est définie dans la bande $\langle -2, -1\rangle$ et y vaut

$$\Theta_3^\star(s) = -2(s+1)\Gamma(s)\frac{p^{-s}+q^{-s}}{(1-(p^{-s}+q^{-s}))^2}.$$

Nous sommes intéressés par le développement asymptotique de $\Theta_3(z)$ au voisinage de l'infini jusqu'au termes linéaires. Dans l'intervalle $]-1, 0]$, la transformée de Mellin a

– un pôle double en -1 en lequel la partie singulière vaut

$$\frac{2}{(h(s+1))^2} - \frac{2}{h^2(s+1)}\left[(\gamma - 1 - h + \frac{h_2}{h})\right],$$

où h est l'entropie de la source et $h_2 := p^2\log p + q^2\log q$, la deuxième entropie.

– dans le cas d'une source périodique, on a aussi des pôles imaginaires sur l'axe $\Re = -1$. Les parties singulières de la transformée de Mellin en les $-1 + 2ik\pi/(\log p - \log q) = -1 + i\alpha_k$, pour $k \in \mathbb{Z}^\star$ sont de la forme

$$-\frac{2i\alpha_k\Gamma(-1+i\alpha_k)}{h^2(s+1-i\alpha_k)^2} - \frac{2}{h^2(s+1-\alpha_k)}\Big[\Gamma(-1+i\alpha_k) + i\alpha_k\Gamma'(-1+i\alpha_k) + i\alpha_k h\Gamma(-1+i\alpha_k) - \frac{h_2 i\alpha_k\Gamma(-1+i\alpha_k)}{h}\Big].$$

On passe rapidement sur les détails de la technique de la transformation de Mellin pour donner le résultat qui nous intéresse.

Lemme 21 *Le comportement asymptotique au voisinage de l'infini de $\Theta_3(z)$ pour une source périodique vaut*

$$\frac{2z\log z}{h^2}\left[1 - \sum_{k\in\mathbb{Z}^\star} i\alpha_k\Gamma(-1+i\alpha_k)z^{-i\alpha_k}\right] + \frac{2z}{h^2}\Big[\left(\gamma - 1 - h + \frac{h_2}{h}\right) + \sum_{k\in\mathbb{Z}^\star}\left(\Gamma(-1+i\alpha_k) + i\alpha_k\Gamma'(-1+i\alpha_k) + i\alpha_k h\Gamma(-1+i\alpha_k) - \frac{h_2 i\alpha_k\Gamma(-1+i\alpha_k)}{h^2}\right)z^{-i\alpha_k}\Big] + o(z), \tag{3.15}$$

et pour une source apériodique

$$\frac{2z\log z}{h^2} + \frac{2z}{h^2}\left(\gamma - 1 - h + \frac{h_2}{h}\right) + o(z). \tag{3.16}$$

Nous venons d'obtenir le comportement asymptotique des sommes $\Theta_2(z)$ et $\Theta_3(z)$. Pour conclure sur le comportement asymptotique de la variance $\mathbb{V}(S)$ de la taille d'un trie dans un modèle de source sans mémoire et de Poisson sur le nombre de textes, il nous faut le comportement asymptotique du carré de l'espérance de la taille $\mathbb{E}^2(S)$.

3.2.5 Contribution asymptotique de $\mathbb{E}^2(S)$

La somme $\Theta_1(z) = \sum_{w\in\mathcal{A}^\star} Q_w(z)$ a un comportement asymptotique en

$$\frac{z}{h}\left(1 - \sum_{k\in\mathbb{Z}^\star} i\alpha_k\Gamma(-1+i\alpha_k)z^{-i\alpha_k}\right) + o(z).$$

quand z tend vers l'infini (cf. chapitre 1). Le développement asymptotique de la somme $\Theta_1(z)$ est étendu aux termes plus grands que la constante et il découle que le comportement asymptotique du carré de l'espérance de la taille vaut

$$\left(\frac{z}{h}\left(1 - \sum_{k\in\mathbb{Z}^\star} i\alpha_k\Gamma(-1+i\alpha_k)\right) - \sum_{j\in\mathcal{J}}\frac{(\beta_j+1)\Gamma(\beta_j)}{H_{\beta_j}}z^{-\beta_j} + O(1)\right)^2 \tag{3.17}$$

où les β_j pour $j \in \mathcal{J}$ sont les zéros simples de $1 - (p^{-s} + q^{-s})$ dans la bande $\langle -1, 0\rangle$ et $H_{\beta_j} = -p^{-\beta_j}\log p - q^{-\beta_j}\log q$. Pour éviter des formules trop lourdes, on restreint encore les calculs au cas d'une source apériodique. Cela signifie que l'on a qu'un seul pôle sur l'axe $\Re = -1$: celui

en $s=-1$ et aucun en $\Re=0$ d'après le lemme 19. Dès lors, le comportement asymptotique de $\mathbb{E}^2(S)$ vaut

$$\begin{aligned}&\left(\frac{z}{h}-\sum_{j\in\mathcal{J}}\frac{(\beta_j+1)\Gamma(\beta_j)}{H_{\beta_j}}z^{-\beta_j}+o(1)\right)^2\\&=\frac{z^2}{h^2}+2\sum_{j\in\mathcal{J}}\frac{(\beta_j+1)\Gamma(\beta_j)}{hH_{\beta_j}}z^{1-\beta_j}+\sum_{(j,k)\in\mathcal{K}}\frac{(\beta_j+1)(\beta_k+1)\Gamma(\beta_j)\Gamma(\beta_k)}{H_{\beta_j}H_{\beta_k}}z^{-\beta_j-\beta_k}+o(z),\end{aligned}\tag{3.18}$$

où $\mathcal{K}$ est l'ensemble des paires d'éléments de $\mathcal{J}$ qui vérifient $\Re(\beta_j+\beta_k)\leq -1$.

3.2.6 Asymptotique de la variance de la taille d'un trie

Dans les paragraphes précédents, nous avons obtenu les différentes contributions au comportement asymptotique de la variance de la taille d'un trie ($\Theta_1(z)$, $\Theta_2(z)$, $\Theta_3(z)$ et $\mathbb{E}^2_{\mathcal{P}(z)}(S)$). On récolte séparément les contributions selon le coefficient en z.

- Les sommes $\Theta_3(z)$ et $\mathbb{E}^2(S)$ apportent des termes en z^2 au comportement asymptotique de $\mathbb{V}(T)$:
$$\frac{2pq}{h^2(1-(p^2+q^2))}-\frac{1}{h^2}=0.$$
- Pour chaque terme en $z^{1-\beta_j}$, avec $j\in\mathcal{J}$, on a des coefficients (à un facteur multiplicatif près)
$$-2\frac{qp^{-\beta_j}+pq^{-\beta_j}}{1-(p^{1-\beta_j}+q^{1-\beta_j})}+2=0,$$
puisque $(p^{-\beta_j}+q^{-\beta_j})(p+q)=1$.
- Chaque terme en $z^{-\beta_j-\beta_k}$ avec $(j,k)\in\mathcal{K}$ contribue (à un facteur multiplicatif près)
$$2\frac{p^{-\beta_j}q^{-\beta_k}}{1-(p^{-\beta_j-\beta_k}+q^{-\beta_j-\beta_k})}-1+2\frac{q^{-\beta_j}p^{-\beta_k}}{1-(p^{-\beta_j-\beta_k}+q^{-\beta_j-\beta_k})}-1=0,$$
en se servant de $(p^{-\beta_j}+q^{-\beta_j})(p^{-\beta_k}+q^{-\beta_k})=1$ et de la symétrie de l'ensemble $\mathcal{K}$ sur lequel on somme. Parmi ces termes on peut trouver des termes en $\beta_j+\beta_k=-1$, mais puisqu'ils s'annulent, il n'est pas nécessaire de les faire apparaître ailleurs.
- Pour les termes en $z\log z$, on a une contribution du comportement asymptotique des termes des sommes $\Theta_2(z)$ et $\Theta_3(z)$:
$$\frac{2}{h^2}-\frac{2}{h^2}=0.$$

Théorème *Le comportement asymptotique de la variance de la taille d'un trie sous un modèle de Poisson de paramètre z et pour une source biaisée sans mémoire (p,q) est linéaire*

$$\mathbb{V}(S)=O(z).$$

3.3 Variance de la longueur de cheminement des tries

Dans cette section nous montrons le comportement asymptotique de la variance de la longueur de cheminement d'un trie dans un modèle de source sans mémoire pour la génération des textes et sous un modèle de Poisson de paramètre z pour le nombre de textes ainsi que sous un modèle de Bernoulli.

La longueur de cheminement L d'un trie a été définie à la page 20 du chapitre 1. Le comportement asymptotique du carré de l'espérance est le carré du comportement asymptotique de l'espérance, or nous avons obtenu au chapitre 1 un développement asymptotique complet de l'espérance de la longueur de cheminement d'un trie. Nous savons ainsi obtenir le comportement asymptotique du carré de la taille. Le point le plus délicat est de trouver le comportement asymptotique de la moyenne de L^2. Comme dans la section précédente, les paires de motifs sont décomposées en deux ensembles selon que ces motifs sont préfixes l'un de l'autre ou non. Nous obtenons

Théorème 6 *Le comportement asymptotique de la variance de la longueur de cheminement d'un trie construit sur des textes engendrés par un modèle de source sans mémoire (p, q) est, pour un modèle de Poisson de paramètre z*

$$\begin{aligned}\mathbb{V}_{\mathcal{P}(z)}(L) =& \frac{1}{h^2} z \log^2 z + K_5 z \log z + O(z), \ \textit{et pour un modèle de Bernoulli,}\\ \mathbb{V}_n(L) =& K_6 n \log n + O(n), \ \textit{où}\\ K_5 :=& \frac{2h_2}{h^3} + \frac{2(1+\gamma)}{h^2} - \frac{1}{h}, \quad K_6 := \frac{h_2 - h^2}{h^3} \ \textit{et} \quad h_2 := p\log^2 p + q \log^2 q.\end{aligned}$$

Dans les premières sections, nous trouvons des formules explicites pour les différentes contributions à l'espérance du carré de la longueur de cheminement sous le modèle de Poisson de paramètre z. La technique de la transformation de Mellin nous sert ensuite à obtenir le comportement asymptotique de chacune des sommes qui entrent en jeu.

Le carré de la longueur de cheminement s'écrit

$$\begin{aligned}L^2 &= \sum_{w\in\mathcal{A}^\star} N_w^2 [\![N_w \geq 2]\!] + \sum_{(w,w')\in(\mathcal{A}^\star)^2:\ w\neq w'} N_w N_{w'} [\![N_w \geq 2]\!][\![N_{w'} \geq 2]\!]\\ &= \sum_{w\in\mathcal{A}^\star} N_w^2 - [\![N_w = 1]\!]\\ &\quad + \sum_{(w,w')\in(\mathcal{A}^\star)^2:\ w\neq w'} N_w N_{w'} - N_{w'}[\![N_w = 1]\!] - N_w[\![N_{w'} = 1]\!] + [\![N_w = 1 \cap N_{w'} = 1]\!]\end{aligned}$$

Il y a quatre espérances à calculer pour les deux ensembles de motifs : cousins ou préfixes ce qui fait 8 cas à étudier. À cela s'ajoute le comportement asymptotique de l'espérance de la somme

$$\Theta_1(z) := \sum_{w\in\mathcal{A}^\star} \mathbb{E}(N_w^2) - \mathbb{P}(N_w = 1). \tag{3.19}$$

Pour la seconde somme, on utilise le même découpage des motifs que dans le cas de la taille : soit les deux motifs sont des cousins, soit l'un des deux est préfixe de l'autre. Les notations sont aussi analogues : la somme sur les cousins est notée $\Theta_2(z)$ et la somme sur les préfixes $\Theta_3(z)$. Dans les prochaines sections, les espérances $\mathbb{E}(N_w N_{w'})$, $\mathbb{E}(N_{w'}[\![N_w = 1]\!])$, $\mathbb{E}(N_w[\![N_{w'} = 1]\!])$ et

$\mathbb{E}([\![N_w = 1 \cap N_{w'} = 1]\!]) = \mathbb{P}(N_w = 1 \text{ et } N_{w'} = 1)$ sont exprimées dans le cas où les deux motifs w et w' sont cousins et dans le cas où un motif est préfixe d'un autre. Les techniques de la transformation de Mellin servent ensuite à obtenir le comportement asymptotique des sommes $\Theta_2(z)$ et $\Theta_3(z)$ quand z tend vers l'infini. Pour le comportement asymptotique de $\Theta_2(z)$, nous appliquons la méthode en « zigzag » comme dans la section précédente.

3.3.1 Expression de la contribution des cousins

Dans cette section, les motifs w et w' sont cousins, c'est-à-dire que $u = w \wedge w'$ le plus long prefixe commun à w et w', n'est ni w, ni w'. On obtient les expressions de

$$\mathbb{E}(N_w N_{w'}),\ \mathbb{E}(N_w [\![N_{w'} = 1]\!]) \text{ et } \mathbb{P}(N_w = 1 \cap N_{w'} = 1).$$

L'expression de $\mathbb{E}(N_{w'}[\![N_w = 1]\!])$ est obtenue à partir de $\mathbb{E}(N_w[\![N_{w'} = 1]\!])$ en utilisant la symétrie entre les motifs w et w'.

➜ $\mathbb{E}(N_w N_{w'})$

La formule pour $\mathbb{E}(N_w N_{w'})$ est obtenue en utilisant la probabilité conditionnelle d'avoir exactement n chaînes qui commencent par le motif u.

$$\begin{aligned}
\mathbb{E}(N_w N_{w'}) =& \sum_{k,k'} kk' \mathbb{P}(N_w = k \text{ et } N_{w'} = k') \\
=& \sum_{k,k'} \sum_{n \geq k+k'} kk' \mathbb{P}(N_w = k \text{ et } N_{w'} = k' \mid N_u = n) \mathbb{P}(N_u = n) \\
=& \sum_{k,k'} \sum_{n \geq k+k'} kk' \binom{n}{k} p_{w|u}^k \binom{n-k}{k'} p_{w'|u}^{k'} (1 - p_{w|u} - p_{w'|u})^{n-(k+k')} \frac{(zp_u)^n}{n!} \exp(-zp_u) \\
=& zp_u p_{w|u} zp_u p_{w'|u}.
\end{aligned}$$

Dans le cas de textes engendrés par une source sans mémoire, l'espérance s'exprime

$$\mathbb{E}(N_w N_{w'}) = zp_w zp_{w'}. \tag{3.20}$$

➜ $\mathbb{E}(N_w[\![N_{w'} = 1]\!])$

Comme pour la taille, on conditionne par le nombre de mots commençant par u dans l'ensemble X sur lequel est formé le trie. Le problème revient à choisir parmi les n mots qui commencent par u, k mots qui commencent par w, un mot parmi les $n-k$ restants qui commence par w' et tous les $n-k-1$ autres qui commencent par u mais ni par w, ni par w'. La loi de

Poisson donne une expression des probabilités qui entrent en jeu et ainsi on écrit

$$\begin{aligned}
\mathbb{E}(N_w[\![N_{w'}=1]\!]) &= \sum_{k>0} k\mathbb{P}(N_w=k\cap N_{w'}=1)\\
&= \sum_{k>0} k \sum_{n\geq k+1} \mathbb{P}(N_w=k\cap N_{w'}=1\,|\,N_u=n)\mathbb{P}(N_u=n)\\
&= \sum_{k>0} k \sum_{n\geq k+1} \binom{n}{k} p_{w|u}^k \binom{n-k}{1} p_{w'|u}\Big(1-p_{w|u}-p_{w'|u}\Big)^{n-k-1} \frac{(zp_u)^n\exp(-zp_u)}{n!}\\
&= (zp_u)^2\exp(-zp_u)p_{w|u}p_{w'|u}\sum_{k>0}\frac{(zp_up_{w|u})^{k-1}}{(k-1)!}\sum_{n\geq k+1}\frac{\Big(zp_u(1-p_{w|u}-p_{w'|u})\Big)^{n-k-1}}{(n-k-1)!}\\
&= (zp_u)^2\exp(-zp_u)p_{w|u}p_{w'|u}\exp(zp_up_{w|u})\exp(zp_u(1-p_{w|u}-p_{w'|u}))\\
&= (zp_u)^2p_{w|u}p_{w'|u}\exp(-zp_up_{w'|u}).
\end{aligned} \tag{3.21}$$

Dans le cas particulier des sources sans mémoire, le résultat s'écrit

$$\mathbb{E}(N_w[\![N_{w'}=1]\!]) = z^2p_wp_{w'}\exp(-zp_{w'}). \tag{3.22}$$

➜ $\mathbb{P}(N_w=1\cap N_{w'}=1)$

On détermine la formule génerale de la probabilité $\mathbb{P}(N_w=1\cap N_{w'}=1)$ sous un modèle de Poisson pour le nombre de chaînes dans l'ensemble de base. La probabilité conditionnelle qui nous intéresse est celle de l'événement : $\{N_w=1\cap N_{w'}=1\}$ sachant $N_u=n$. Soit en reformulant : si on sait qu'exactement n mots de l'ensemble de base commencent par u alors un seul d'entre eux commence par w, un seul par w' (et ce n'est pas le même mot qui vérifie ces deux propriétés par définition de « cousin ») et les $n-2$ autres ne commencent ni par w, ni par w' mais quand même par u. Cela se transcrit par les équations :

$$\mathbb{P}(N_w=1\cap N_{w'}=1\,|\,N_u=n) = \binom{n}{1}p_{w|u}\binom{n-1}{1}p_{w'|u}(1-p_{w|u}-p_{w'|u})^{n-2}$$

La probabilité d'avoir exactement une chaîne commençant par w et exactement une chaîne commençant par w' vaut

$$\mathbb{P}(N_w=1\cap N_{w'}=1) = (zp_u)^2p_{w|u}p_{w'|u}\exp(-zp_up_{w|u})\exp(-zp_up_{w'|u}). \tag{3.23}$$

Dans le cas particulier d'une source sans mémoire, le résultat s'écrit

$$\mathbb{P}(N_w=1\cap N_{w'}=1) = (zp_w\exp(-zp_w))(zp_{w'}\exp(-zp_{w'})).$$

Nous venons d'obtenir les expressions des probabilités nécessaires pour reformuler la somme

$$\Theta_2(z) = \sum_{(w,w')\text{ cousins}} zp_w(1-\exp(-zp_w))zp_{w'}(1-\exp(-zp_{w'})). \tag{3.24}$$

De plus, on a déjà vu dans le cas de la taille que deux motifs cousins w et w' se récrivent en fonction de leur plus long préfixe commun u. Les lettres $w_{|u|+1}$ et $w'_{|u|+1}$ doivent forcément différer. Les motifs peuvent s'écrire $w=u.\mathtt{1}.v$ et $w'=u.\mathtt{0}.v'$ ou $w=u.\mathtt{1}.v$ et $w'=u.\mathtt{0}.v'$. La symétrie de l'ensemble en w et w' permet de décider que le motif w « vire à gauche » après u, c'est-à-dire s'écrit $u\mathtt{0}v$ et que le motif w' « vire à droite » et donc s'écrit $u\mathtt{1}v'$. Pour tenir compte de ce choix on multiplie la somme par 2. La somme $\Theta_2(z)$ sur les cousins se reformule par

Lemme 22 *Sous un modèle de source sans mémoire sur la génération des textes et un modèle de Poisson de paramètre z sur le nombre de textes, on a*

$$\Theta_2(z) = 2 \sum_{(u,v,v')\in(\mathcal{A}^\star)^3} zp_{u0v}(1-\exp(-zp_{u0v}))zp_{u1v'}(1-\exp(-zp_{u1v'})).$$

3.3.2 Expression de la contribution des préfixes

Dans cette section, on traite le cas où un des deux motifs est préfixe de l'autre. La symétrie de la somme nous permet de prendre arbitrairement w préfixe de w'. La somme sera multipliée par 2 pour tenir compte de cette rupture de la symétrie originale. Nous sommes intéressés par les expressions de

$$\mathbb{E}(N_w N_{w'}),\ \mathbb{E}(N_w[\![N_{w'}=1]\!]),\ \mathbb{E}(N_{w'}[\![N_w=1]\!]) \text{ et } \mathbb{P}(N_w=1\cap N_{w'}=1).$$

➜ $\mathbb{E}(N_w N_{w'})$

La détermination de $\mathbb{E}(N_w N_{w'})$ est plus compliquée pour w et w' préfixes que dans le cas de motifs cousins. On évalue, *bis repetita placent*, la probabilité conditionnelle quand on sait que k mots commencent par w, c'est-à-dire on se place dans un modèle de Bernoulli pour les chaînes commençant par le préfixe w et les résultats sur la loi de Poisson (N_w suit une loi de Poisson) permettent d'obtenir le résultat. Le principe de la preuve repose sur l'écriture de l'espérance sous la forme

$$\mathbb{E}(N_w N_{w'}) = \sum_{k,k'} kk'\mathbb{P}(N_w=k\cap N_{w'}=k') = \sum_{k,k'} kk'\mathbb{P}(N_{w'}=k' \mid N_w=k)\mathbb{P}(N_w=k).$$

Il nous faut calculer

$$\begin{aligned}\mathbb{E}(N_w N_{w'}) &= \sum_{k,k'} kk' \binom{k}{k'} p_{w'|w}^{k'}(1-p_{w'|w})^{k-k'} \frac{(zp_w)^k}{k!}\exp(-zp_w)\\ &= \exp(-zp_w)\sum_{j\geq 0}\frac{1}{(j-1)!}(zp_w(1-p_{w'|w}))^j \left[\sum_{k'\geq 0}\frac{(zp_w p_{w'|w})^{k'}}{(k'-1)!}\right]\\ &\quad + \frac{1}{j!}(zp_w(1-p_{w'|w}))^j \left[\sum_{k'\geq 0}\frac{k'(zp_w p_{w'|w})^{k'}}{(k'-1)!}\right],\end{aligned}$$

avec un changement de variables $j=k-k'$. Après un monceau de calculs, nous aboutissons à

$$\mathbb{E}(N_w N_{w'}) = zp_w p_{w'|w}(zp_w+1), \tag{3.25}$$

ce qui pour une source sans mémoire se traduit par

$$\mathbb{E}(N_w N_{w'}) = zp_{w'}(zp_w+1). \tag{3.26}$$

➜ $\mathbb{P}(N_w = 1 \cap N_{w'} = 1)$

La probabilité $\mathbb{P}(N_w = 1 \cap N_{w'} = 1)$ se traduit comme la probabilité qu'il y ait un mot unique qui commence par w et un unique mot par w'. Or w est préfixe de w' donc il s'agit de trouver la probabilité que le seul mot de l'ensemble de base qui commence par w commence aussi par w'. Soit

$$\begin{aligned}\mathbb{P}(N_w = 1 \cap N_{w'} = 1) &= \mathbb{P}(N_{w'} = 1 \,|\, N_w = 1)\mathbb{P}(N_w = 1)\\ &= zp_w p_{w'|w} \exp(-zp_w),\end{aligned}$$

et dans le cas d'une source sans mémoire, on a la formule

$$\mathbb{P}(N_w = 1 \cap N_{w'} = 1) = zp_{w'} \exp(-zp_w). \tag{3.27}$$

➜ $\mathbb{E}(N_{w'}[\![N_w = 1]\!])$

Il n'y a plus de symétrie entre w et w', il nous faut regarder séparément $\mathbb{E}(N_{w'}[\![N_w = 1]\!])$ et $\mathbb{E}(N_w[\![N_{w'} = 1]\!])$. Cependant une remarque nous simplifie grandement le calcul de $\mathbb{E}(N_{w'}[\![N_w = 1]\!])$: puisqu'il n'y a qu'une seule chaîne qui commence par w, il ne peut y en avoir au plus qu'une qui commence par w'. D'où

$$\mathbb{E}(N_{w'}[\![N_w = 1]\!]) = \sum_{k>0} k\mathbb{P}(N_w = 1 \cap N_{w'} = k) = \mathbb{P}(N_w = 1 \cap N_{w'} = 1). \tag{3.28}$$

➜ $\mathbb{E}(N_w[\![N_{w'} = 1]\!])$

On a recours à la même méthode que préalablement : le conditionnement par le nombre de chaînes qui commencent par w, soit :

$$\begin{aligned}\mathbb{E}(N_w[\![N_{w'} = 1]\!]) &= \sum_{k>0} k\mathbb{P}(N_w = k \cap N_{w'} = 1)\\ &= \sum_{k>0} k\mathbb{P}(N_{w'} = 1 \,|\, N_w = k)\mathbb{P}(N_w = k)\\ &= \sum_{k\geq 1} k\binom{k}{1} p_{w'|w}(1 - p_{w'|w})^{k-1} \exp(-zp_w)\frac{(zp_w)^k}{k!}\\ &= zp_w p_{w'|w} \exp(-zp_w p_{w'|w})[zp_w(1 - p_{w'|w}) + 1],\end{aligned}$$

et, en particulier, pour une source sans mémoire,

$$\mathbb{E}(N_w[\![N_{w'} = 1]\!]) = zp_{w'} \exp(-zp_{w'})(zp_w - zp_{w'} + 1). \tag{3.29}$$

Ces résultats sur les probabilités, nous permettent d'exprimer la somme $\Theta_3(z)$

$$\Theta_3(z) = 2 \sum_{(w,w')\colon w \sqsubset w'} zp_{w'}((zp_w + 1) - \exp(-zp_{w'})(zp_w - zp_{w'} + 1)).$$

De plus, le fait que w soit préfixe de w' permet d'écrire $w' = w.u$ avec un motif u appartenant à $\mathcal{A}^+$. Ainsi

Lemme 23 *Dans un modèle de source sans mémoire pour la génération des textes et un modèle de Poisson de paramètre z sur le nombre de chaînes de l'ensemble de base, la contribution des motifs préfixes l'un de l'autre à l'espérance du carré de la longueur de cheminement s'exprime*

$$\Theta_3(z) = 2 \sum_{w\in\mathcal{A}^\star} \sum_{u\in\mathcal{A}^+} zp_wp_u((zp_w+1) - \exp(-zp_wp_u)(zp_w - zp_wp_u + 1)).$$

Nous venons d'obtenir les expressions des sommes $\Theta_2(z)$ et $\Theta_3(z)$. Dans les sections suivantes nous obtenons le comportement asymptotique des trois somme $\Theta_i(z)$ quand z tend vers l'infini à l'aide de la transformation de Mellin.

3.3.3 Contribution asymptotique de la somme $\Theta_1(z)$

Nous exprimons la somme $\Theta_1(z)$ avant de déterminer le comportement asymptotique quadratique de cette somme quand z tend vers l'infini.

La variable aléatoire N_w suit une loi de Poisson, donc

$$\mathbb{P}(N_w = 1) = zp_w \exp(-zp_w) \text{ et } \mathbb{E}(N_w^2) = \sum_{k\geq 0} k^2\mathbb{P}(N_w = k) = zp_w(1 + zp_w).$$

La somme $\Theta_1(z)$ s'exprime de façon simple

$$\Theta_1(z) = \sum_{w\in\mathcal{A}^\star} f(zp_w), \text{ où } f(z) = z(1 + z - \exp(-z)). \tag{3.30}$$

Au voisinage de zéro comme au voisinage de l'infini, la fonction $f(z)$ se comporte en $O(z^2)$ donc la bande fondamentale de la transformée de Mellin de cette fonction est vide. Pour résoudre ce problème, on scinde la somme en deux :

$$\Theta_1(z) = \sum_w (zp_w)^2 + \sum_w zp_w(1 - \exp(-zp_w)),$$

les deux nouvelles sommes convergent donc cette opération est bien permise. La première $\Phi_1(z) := \sum_w (zp_w)^2$ est déterminée explicitement, pour la seconde $\Phi_2(z) := \sum_w zp_w(1 - \exp(-zp_w))$ (déjà rencontrée dans le chapitre 1) on utilise les techniques de base de la théorie de la transformation de Mellin. La fonction de base dans l'étude de $\Phi_2(z)$ est nommée $g(z) = z(1 - \exp(-z))$, sa bande fondamentale est $\langle -2, -1\rangle$ et la transformée de Mellin de la somme $G(z) := \sum_w g(zp_w)$ vaut $\sum_w p_w^{-s} g^\star(s)$. La bande de convergence de la série de Dirichlet est $\langle -\infty, -1\rangle$, la transformée vaut $g^\star(s) = -s\Gamma(s)$ et ainsi dans la bande $\langle -2, -1\rangle$ on a

$$G^\star(s) = -s\Gamma(s)\frac{1}{1 - (p^{-s} + q^{-s})}.$$

Seuls les pôles situés sur l'axe $\Re = -1$ nous intéressent pour connaître le comportement asymptotique jusqu'en $o(z)$. Il y a un pôle double en $s = -1$ et des pôles simples en $-1 + 2ik\pi/(\log p - \log q) =: -1 + i\alpha_k$ (si la source est périodique *i.e.* si $\log p/\log q$ est rationnel) avec k entier non nul. Après une application du théorème de Mellin inverse, on obtient le comportement asymptotique de la somme simple $\Theta_1(z)$:

Lemme 24 *Dans un modèle de source sans mémoire (p, q) et un nombre de mots de l'ensemble de base qui suit une loi de Poisson de paramètre z, la somme $\Theta_1(z)$ se comporte asymptotiquement au voisinage de l'infini en*

$$\frac{z^2}{1-(p^2+q^2)} + \frac{1}{h} z \log z + \frac{z}{h}\left(\frac{h_2}{2h} + \gamma - \sum_{k\in\mathbb{Z}^\star} \Gamma(-1+i\alpha_k)(i\alpha_k - 1) z^{-i\alpha_k}\right) + o(z) \tag{3.31}$$

si la source est périodique et

$$\frac{z^2}{1-(p^2+q^2)} + \frac{1}{h} z \log z + \frac{z}{h}\left(\frac{h_2}{2h} + \gamma\right) + o(z) \text{ sinon.} \tag{3.32}$$

La constante $h = -p\log p - q\log q$ est l'entropie de la source, $h_2 = p\log^2 p + q\log^2 q$ est la deuxième entropie de la source et $\alpha_k = 2k\pi/(\log p - \log q)$ pour tout entier k.

3.3.4 Contribution asymptotique des cousins

La contribution des motifs cousins est donnée par le lemme 22

$$\Theta_2(z) = 2 \sum_{(u,v,v')\in(\mathcal{A}^\star)^3} z p_{u0v}(1-\exp(-z p_{u0v})) z p_{u1v'}(1-\exp(-z p_{u1v'})).$$

Nous déterminons le comportement asymptotique de cette somme en utilisant la méthode en « zigzag » de la transformée de Mellin.

On reprend la définition $g(z) := z(1-\exp(-z))$ du paragraphe précédent. La première étape de l'application en zigzag consiste à calculer le comportement asymptotique de $G_p(z) := \sum_v g(zpp_v)$ (ou de manière équivalente $G_q(z) := \sum_v g(zqp_v)$) quand z tend vers l'infini en se servant du théorème de Mellin inverse. À l'aide des résultats du paragraphe précédent, on a

$$G_p^\star(s) = -s\Gamma(s) p^{-s} \frac{1}{1-(p^{-s}+q^{-s})}.$$

La transformée de Mellin a un pôle double en -1 et des pôles simples en les $-1+i\alpha_k$ (où $\alpha_k := 2k\pi/(\log p - \log q)$ avec $k \in \mathbb{Z}^\star$) si la source est périodique. Il y a aussi tous les zéros de l'équation $p^{-s} + q^{-s} = 1$ qui sont dans la bande $\langle -1, 2\rangle$. Tous ces pôles de la transformée de Mellin sont simples ([FFH86]), mais leur localisation précise n'est pas connue. Ils sont notés β_j pour $j \in \mathcal{L}$, et leur résidu est γ_j. En zéro il y a une singularité effaçable de la fonction Gamma. Dans le cas d'une source périodique, le comportement asymptotique est

$$\frac{p}{h} z\log z + \frac{zp}{h}\left(\log p + \gamma + \frac{h_2}{2h} - \sum_{k\in\mathbb{Z}^\star} \Gamma(-1+i\alpha_k) p^{-i\alpha_k} (i\alpha_k - 1) z^{-\alpha_k}\right) - \sum_{j\in\mathcal{L}} \beta_j \Gamma(\beta_j) p^{-\beta_j} \gamma_j z^{-\beta_j} + O(z^{-2}),$$

et dans le cas apériodique, le comportement asymptotique est

$$\frac{p}{h} z\log z + \frac{zp}{h}\left(\log p + \gamma + \frac{h_2}{2h}\right) - \sum_{j\in\mathcal{J}} \beta_j \Gamma(\beta_j) p^{-\beta_j} \gamma_j z^{-\beta_j} + O(z^{-2}).$$

Le comportement asymptotique de $\sum_{v,v'\in\mathcal{A}^\star} g(zpp_v) g(zqp_{v'})$ est le produit des comportements asymptotiques de $G_p(z)$ et $G_q(z)$. Il permet de determiner le développement singulier de la transformée de Mellin de $G_p(z)G_q(z)$. Puisque le comportement asymptotique dominant du

produit en l'infini est en $z^2\log^2 z$, le bord droit de la bande fondamentale de $(G_pG_q)^\star$ est -2. La bande fondamentale n'est pas vide puisque le comportement asymptotique du produit G_pG_q au voisinage de zéro se comporte en z^4. Donc le bord gauche de la bande fondamentale est -4. Dans la bande fondamentale $\langle -4,-2\rangle$ la transformée de Mellin vaut

$$\Theta_2^\star(s) = 2(G_pG_q)^\star(s)\sum_{u\in\mathcal{A}^\star} p_u^{-s} = 2(G_pG_q)^\star(s)\frac{1}{1-(p^{-s}+q^{-s})}.$$

Dans la suite de cette section, nous ne regardons que le cas d'une source apériodique pour que le lecteur ne soit pas noyé sous les calculs déjà pesants. Le produit G_pG_q se comporte asymptotiquement en

$$\begin{aligned}
&\left(\frac{p}{h}z\log z + \frac{zp}{h}\left(\log p+\gamma+\frac{h_2}{2h}\right) - \sum_{j\in\mathcal{L}}\beta_j\Gamma(\beta_j)p^{-\beta_j}\gamma_j z^{-\beta_j} + O(z^{-2})\right)\times\\
&\qquad\left(\frac{q}{h}z\log z + \frac{zq}{h}\left(\log q+\gamma+\frac{h_2}{2h}\right) - \sum_{j\in\mathcal{L}}\beta_j\Gamma(\beta_j)q^{-\beta_j}\gamma_j z^{-\beta_j} + O(z^{-2})\right)\\
&= \frac{pq(z\log z)^2}{h^2} + \frac{pqz^2\log z}{h^2}(a_p+a_q) + \frac{pqz^2(a_pa_q)}{h^2} - \frac{z\log z}{h}\sum_{j\in\mathcal{G}}(qb_{p,j}+pb_{q,j})z^{-\beta_j}\\
&\qquad - \frac{z}{h}\sum_{j\in\mathcal{G}}(qb_{p,j}a_q+pb_{q,j}a_p)z^{-\beta_j} + \sum_{(i,j)\in\mathcal{M}} z^{-\beta_j-\beta_i}b_{p,i}b_{q,j} + o(z),
\end{aligned}$$

où

$$a_p := \log p+\gamma+h_2/2h,\quad b_{p,j} := \beta_j\Gamma(\beta_j)p^{-\beta_j}\gamma_j,$$

$\mathcal{G}$ est l'ensemble des indices i pour lesquels $\Re(\beta_i)\le 0$ et $\mathcal{M}$ est l'ensemble des paires d'indices (i,j) telles que $\Re(\beta_i+\beta_j)\le -1$. Pour nos besoins, nous cherchons juste un développement asymptotique du produit G_pG_q jusqu'au terme en $o(z)$. Nous sommes dans le cas d'une source apériodique, et donc il n'y a pas de pôle vérifiant $\beta_j=0$ (résultat du lemme 19). Ainsi l'ensemble $\mathcal{G}$ est l'ensemble des j tels que β_j soit dans la bande $\langle -1,0\rangle$.

Le développement singulier de la transformée de Mellin $\Theta_2^\star(s)$ est obtenu à partir du développement singulier de $(G_pG_q)^\star$ (*via* une application du théorème de Mellin « direct ») et du développement singulier de la série de Dirichlet associée. Il faut faire particulièrement attention aux produits croisés : les termes en $(s+2)^2$ du développement singulier de la série de Dirichlet se multiplie avec le terme en $(s+2)^{-3}$ du développement singulier de $(G_pG_q)^\star$. En résumé et en omettant temporairement le facteur 2 dans $\Theta_2(z)$, il y a

– en -2, un pôle triple de contribution

$$\begin{aligned}
&\frac{pq}{h^2(1-(p^2+q^2))}\Big(1+(s+2)\frac{-p^2\log p-q^2\log q}{1-(p^2+q^2)}\\
&+(s+2)^2\left(\frac{1}{2}\frac{p^2\log^2 p+q^2\log^2 q}{1-(p^2+q^2)}+\left(\frac{-p^2\log p-q^2\log q}{1-(p^2+q^2)}\right)^2\right)\Big)\left(-\frac{2}{(s+2)^3}+\frac{(a_p+a_q)}{(s+2)^2}-\frac{a_pa_q}{(s+2)}\right).
\end{aligned}$$

– en $-1+\beta_j$, un pôle double pour chaque $\beta_j\in\mathcal{G}$ de contribution

$$\begin{aligned}
&\frac{1}{h(1-(p^{1-\beta_j}+q^{1-\beta_j}))}\left(1+(s+1-\beta_j)\frac{-p^{1-\beta_j}\log p-q^{1-\beta_j}\log q}{1-(p^{1-\beta_j}+q^{1-\beta_j})}\right)\times\\
&\qquad\left(-\frac{qb_{p,j}+pb_{q,j}}{(s+1-\beta_j)^2}+\frac{qb_{p,j}a_q+pb_{q,j}a_p}{(s+1-\beta_j)}\right).
\end{aligned}$$

– en $-\beta_j - \beta_i$, un pôle simple si $\Re(\beta_i + \beta_j) \leq -1$ et $\beta_i + \beta_j \neq -1$ de contribution

$$\frac{b_{p,i} b_{q,j}}{(1 - (p^{-\beta_j - \beta_i} + q^{-\beta_i - \beta_j}))(s - \beta_i - \beta_j)},$$

– et en $\beta_i + \beta_j = -1$, on a un pôle double qui contribue

$$\frac{b_{p,i} b_{q,j}}{h(s+1)^2} + \frac{1}{s+1}\left(-\frac{h_2 b_{p,i} b_{q,j}}{2h^2} - \frac{C}{h}\right).$$

Le développement singulier en -2 a une écriture plus simple en notant que

$$1 - (p^2 + q^2) = (p+q)^2 - p^2 - q^2 = 2pq.$$

On se sert maintenant du théorème de Mellin « inverse » qui nous offre un passage entre le développement singulier de la transformée de Mellin $\Theta_2^\star(s)$ et le comportement asymptotique de la contribution $\Theta_2(z)$ des cousins à la moyenne de la longueur de cheminement :

Lemme 25 *Sous un modèle de source sans mémoire et un nombre de textes de l'ensemble de base du trie qui suit une loi de Poisson de paramètre z, la somme $\Theta_2(z)$ se comporte asymptotiquement quand z tend vers l'infini en*

$$\begin{aligned}
&\frac{1}{h^2} z^2 \log^2 z + \frac{1}{h^2}\left(a_p + a_q + \frac{p^2 \log p + q^2 \log q}{pq}\right) z^2 \log z \\
&+ \frac{1}{h_2}\left(a_p a_q - (a_p + a_q)\frac{p^2 \log p + q^2 \log q}{2pq} - 2\left(\frac{1}{2}\frac{p^2 \log^2 p + q^2 \log^2 q}{2pq} + \left(\frac{p^2 \log p + q^2 \log q}{2pq}\right)^2\right)\right) z^2 \\
&+ \sum_{j \in \mathcal{G}} \frac{2}{h(1 - (p^{1-\beta_j} + q^{1-\beta_j}))}\Big(- (q b_{p,j} + p b_{q,j}) z^{1-\beta_j} \log z \\
&+ \left[-(q b_{p,j} a_q + p b_{q,j} a_p) + \frac{(q b_{p,j} + p b_{q,j}) H_{\beta_j - 1}}{1 - (p^{1-\beta_j} + q^{1-\beta_j})}\right] z^{1-\beta_j}\Big) \\
&- \sum_{(i,j) \in \mathcal{M}:\ \beta_i + \beta_j \neq -1} \frac{2 b_{p,i} b_{q,j}}{(1 - (p^{-\beta_j - \beta_i} + q^{-\beta_i - \beta_j}))} z^{-\beta_i - \beta_j} \\
&+ \frac{2}{h} \sum_{(i,j) \in \mathcal{M}:\ \beta_i + \beta_j = -1} \left(b_{p,i} b_{q,j} z \log z + z\left(\frac{h_2 b_{p,i} b_{q,j}}{2h} + C\right)\right) + O(z).
\end{aligned}$$

3.3.5 Contribution asymptotique des préfixes

La contributions des mots « préfixes » à l'espérance de L^2 vaut (à un facteur 2 près dû à la symétrie de la somme)

$$\Delta_3(z) := \frac{1}{2}\Theta_3(z) = \sum_{w \in \mathcal{A}^\star} \sum_{u \in \mathcal{A}^+} z p_w p_u (z p_w + 1 - \exp(-z p_w p_u)(z p_w - z p_w p_u + 1)).$$

Au voisinage de zéro, le terme général de la somme se comporte en $2z^2(p_u p_w)^2$ et au voisinage de l'infini en $z^2 p_w^2 p_u$. La bande fondamentale de la transformée de Mellin du terme général est

donc vide et il faut légèrement modifier le terme général pour que sa transformée de Mellin ait une bande fondamentale non vide. On écrit

$$\Delta_3(z) = \sum_{w\in\mathcal{A}^\star}\sum_{u\in\mathcal{A}^+} zp_wp_u(zp_w+1-\exp(-zp_wp_u)(zp_w-zp_wp_u+1))-2(zp_wp_u)^2+2(zp_wp_u)^2.$$

Le comportement asymptotique de la somme

$$\sum_{w\in\mathcal{A}^\star}\sum_{u\in\mathcal{A}^+}(zp_wp_u)^2 = z^2\sum_{k\geq 0}(p^2+q^2)^k\sum_{j\geq 1}(p^2+q^2)^j = \frac{z^2(p^2+q^2)}{(1-(p^2+q^2))^2}$$

est obtenu sans difficulté.

Le comportement asymptotique de l'autre somme au voisinage de zéro est en z^3 (on a retiré le terme dominant en z^2) et au voisinage de l'infini, le comportement du terme général est en z^2. Le terme général de la somme se récrit à l'aide de deux fonctions

$$zp_wp_u(zp_w+1-\exp(-zp_wp_u)(zp_w-zp_wp_u+1)-2(zp_wp_u)) = \frac{1}{p_u}f_1(zp_wp_u)+f_2(zp_wp_u),$$

où $f_1(z) := z^2(1-\exp(-z))$ et $f_2(z) = z(1-\exp(-z)(1-z)-2z)$. Pour chacune de ces fonctions, la bande fondamentale est $\langle -3,-2\rangle$ et les transformées valent

$$f_1^\star(s) = -s(s+1)\Gamma(s) \text{ et } f_2^\star(s) = s^2\Gamma(s).$$

La transformée de Mellin de la somme de la contribution des préfixes vaut :

$$\begin{aligned}\Delta_3^\star(s) &= \sum_{w\in\mathcal{A}^\star}\sum_{u\in\mathcal{A}^+}(p_up_w)^{-s}\left[\frac{1}{p_u}f_1^\star(s)+f_2^\star(s)\right]\\ &= s\Gamma(s)\frac{1}{1-(p^{-s}+q^{-s})}\left[-(s+1)\frac{p^{-s-1}+q^{-s-1}}{1-(p^{-s-1}+q^{-s-1})}+s\frac{p^{-s}+q^{-s}}{1-(p^{-s}+q^{-s})}\right].\end{aligned}$$

Il y a plusieurs pôles de la transformée à droite de la bande fondamentale : en -2, on a un pôle double, et si la source est périodique, en chaque $-2+\alpha_k$ avec $\alpha_k = 2ik\pi/(\log p-\log q)$ et $k\in\mathbb{Z}^\star$, la transformée a un pôle simple. Il y a aussi tous les zéros (simples d'après [FFH86]) de l'équation $p^{-s}+q^{-s}=1$ qui sont dans la bande $\langle -1,0\rangle$ et par translation, les zéros de $p^{-s-1}+q^{-s-1}=1$ dans la bande $\langle -2,-1\rangle$. Tous ces pôles sont simples puisque la fonction Gamma ne s'y annule pas, et on les note β_j pour $j\in\mathcal{J}$, et le résidu de $(1-(p^{-s}+q^{-s}))^{-1}$ y vaut γ_j. En $s=-1$, la transformée admet un pôle triple et pour les autres pôles sur l'axe en $-1+\alpha_k$ avec $k\in\mathbb{Z}^\star$, la transformée admet des pôles doubles. On regarde le cas d'une source apériodique ce qui nous épargne les calculs fastidieux pour les pôles imaginaires sur les axes $\Re=-2$ et $\Re=-1$.

De plus l'apériodicité de la source garantit qu'on n'ait pas de pôles imaginaires pour $(1-(p^{-s-1}+q^{-s-1}))^{-1}$ avec une partie réelle valant -1. Pour justifier cela, on décale le problème en 0 et on applique le lemme 19.

Il y a trois types de pôles : un pôle double en -2, des pôles simples en les β_j et un pôle triple en -1.

Lemme 26 *Le comportement asymptotique de la contribution Θ_3 des préfixes à l'espérance du carré de la longueur de cheminement d'un trie construit sur des textes engendrés par une source sans mémoire (p,q) et un modèle de Poisson de paramètre z sur le nombre de textes s'écrit*

$$\frac{2z^2\log z}{h(1-(p^2+q^2))}-2K_1z^2-2\sum_{j\in\mathcal{J}}\Gamma(\beta_j+1)K_2\frac{z^{1-\beta_j}}{H_{\beta_j}}+\frac{z\log^2 z}{h^2}+2K_3z\log z+2K_4z+o(z), \quad (3.33)$$

avec les notations

$$
\begin{aligned}
H_{\beta_j} &= -p^{-\beta_j}\log p - q^{-\beta_j}\log q,\\
K_1 &:= \frac{1}{1-(p^2+q^2)}\left(\frac{h+\gamma-h_2/2pq}{h}+\frac{p^2+q^2}{pq}-\frac{p^2\log p+q^2\log q}{2hpq}\right),\\
K_2 &:= \frac{1}{1-(p^{-\beta_j+1}+q^{-\beta_j+1})},\\
K_3 &:= \frac{h_2+(1+\gamma-h)h}{h^3},\\
K_4 &:= \frac{1}{6h^4}(-2pq\log p\log^3 q-2pq\log^3 p\log q+6pq\log^2 p\log^2 q+q^2\log^4 q+p^2\log^4 p)\\
&\quad+\frac{h_2^2}{4h^4}-\frac{-h_2+2h(1+\gamma)-\frac{\pi^2}{6}+\gamma^2-2\gamma}{2h^2}+\frac{(1+\gamma-h)h_2}{h^3}+\frac{2}{h},\ \textit{et}\\
\mathcal{J} &= \{j:\beta_j\ \textit{vérifie}\ p^{-\beta_j}+q^{-\beta_j}=1\ \textit{et}\ -1\le\Re(\beta_j)\le 0\}.
\end{aligned}
$$

3.3.6 Contribution asymptotique de $\mathbb{E}^2(L)$

La variance de la longueur de cheminement dans un trie fait intervenir les trois sommes $\Theta_i(z)$, mais aussi le carré de l'espérance de la longueur de cheminement dans un trie. Dans ce paragraphe, on obtient le comportement asymptotique du carré de l'espérance de la longueur de cheminement.

Le comportement asymptotique de la moyenne de la longueur de cheminement dans un trie sous modèle de Poisson de paramètre z a déjà été déterminé précédemment. On note $\mathcal{H}$ l'ensemble des indices des zéros de $p^{-s}+q^{-s}=1$ situés dans la bande $\langle -1,1\rangle$. Le comportement asymptotique de la longueur de cheminement au voisinage de l'infini vaut

$$
\begin{aligned}
&\left(\frac{z\log z}{h}+\frac{z}{h}\left(\frac{h_2}{2h}+\gamma\right)-\sum_{i\in\mathcal{H}}\frac{\beta_j\Gamma(\beta_j)z^{-\beta_j}}{H_{\beta_j}}+O(z^{-1})\right)^2\\
&=\frac{(z\log z)^2}{h^2}+2\frac{z^2\log z}{h^2}\left(\frac{h_2}{2h}+\gamma\right)+\left(\frac{z}{h}\left(\frac{h_2}{2h}+\gamma\right)\right)^2\\
&-2\sum_{i\in\mathcal{H}}\frac{\beta_j\Gamma(\beta_j)z^{-\beta_j+1}}{hH_{\beta_j}}(\log z+\frac{h_2}{2h}+\gamma)+\sum_{(i,j)\in\mathcal{H}'}\frac{\beta_i\beta_j\Gamma(\beta_i)\Gamma(\beta_j)z^{-\beta_j-\beta_i}}{H_{\beta_j}H_{\beta_i}}+o(z),
\end{aligned}
$$

où $\mathcal{H}'$ est l'ensemble des paires (i,j) avec $-2<\Re(\beta_i+\beta_j)\le -1$.

3.3.7 Asymptotique de la variance de la longueur de cheminement d'un trie

Dans cette section on rassemble les résultats obtenus dans les sections précédentes. La variance de la longueur de cheminement s'écrit

$$\mathbb{V}(L)=\mathbb{E}(L^2)-\mathbb{E}^2(L)=\Theta_1(z)+2\Theta_2(z)+2\Theta_3(z)-\mathbb{E}^2(L).$$

Comme dans le cas de la taille on regarde les coefficients en z les uns après les autres pour collecter l'information asymtotique sur la variance de la longueur de cheminement.

– Pour le coefficient en $(z\log z)^2$ du comportement asymptotique de la variance, on a des contributions de la somme sur les cousins et de $\mathbb{E}^2(P)$

$$\frac{1}{h^2}-\frac{1}{h^2}=0.$$

– Pour le coefficient en $z^2\log z$, on a une contribution

$$\frac{1}{h^2}\left((a_p+a_q)+2\frac{p^2\log p+q^2\log q}{2pq}\right)+\frac{2}{h(1-(p^2+q^2))}-\frac{2}{h^2}\left(\frac{h_2}{2h}+\gamma\right)=0.$$

– Pour le coefficient en z^2, on a une contribution des quatres sommes

$$\begin{aligned}&\frac{1}{1-(p^2+q^2)}-2K_1+4\frac{(p^2+q^2)}{(1-(p^2+q^2))^2}-\frac{1}{h^2}\left(\frac{h_2}{2h}+\gamma\right)^2\\&+\frac{1}{h^2}\left(a_pa_q+(a_p+a_q)\frac{p^2\log p+q^2\log q}{2pq}-2\left(\frac{p^2\log^2 p+q^2\log^2 q}{4pq}+\left(\frac{p^2\log p+q^2\log q}{2pq}\right)^2\right)\right)=0\end{aligned}$$

– Pour le coefficient $z^{1-\beta_j}\log z$ avec $j\in\mathcal{H}$, on a les termes

$$-2\frac{qb_{p,j}+pb_{q,j}}{h(1-(p^{1-\beta_j}+q^{1-\beta_j}))}+2\frac{\beta_j\Gamma(\beta_j)}{hH_{\beta_j}}=2\frac{\beta_j\Gamma(\beta_j)}{hH_{\beta_j}}\left(-\frac{qp^{-\beta_j}+pq^{-\beta_j}}{1-(p^{1-\beta_j}+q^{1-\beta_j})}+1\right)=0.$$

– Pour chaque coefficient $z^{1-\beta_j}$, on a les contributions venant du carré de l'espérance, des motifs cousins et des motifs préfixes :

$$\begin{aligned}-\frac{2\Gamma(\beta_j+1)}{H_{\beta_j}}\Big[&\frac{1}{(1-(p^{1-\beta_j}+q^{1-\beta_j}))}-\left(\gamma+\frac{h_2}{2h}\right)\frac{1}{h}-\frac{(qp^{-\beta_j}+pq^{-\beta_j})H_{\beta_j-1}}{h(1-(p^{1-\beta_j}+q^{1-\beta_j}))^2}\\&+\frac{qp^{-\beta_j}a_q+pq^{-\beta_j}a_p}{h(1-(p^{1-\beta_j}+q^{1-\beta_j}))}\Big]=0.\end{aligned}$$

– Pour le coefficient $z\log z$, le coefficient du comportement asymptotique est formé par des termes provenant des sommes $\Theta_1(z),\Theta_2(z)$ et $\Theta_3(z)$:

$$\begin{aligned}&\frac{1}{h}+\frac{2}{h}\sum_{(i,j)\in\mathcal{M}:\ \beta_i+\beta_j=-1}b_{p,i}b_{q,j}+2K_3\\&=\frac{2h_2}{h^3}+\frac{2(1+\gamma)}{h^2}-\frac{1}{h}+\frac{2}{h}\sum_{(i,j)\in\mathcal{M}:\ \beta_i+\beta_j=-1}\frac{\beta_i\beta_j\Gamma(\beta_i)\Gamma(\beta_j)p^{-\beta_i}q^{-\beta_j}}{H_{\beta_i}H_{\beta_j}}=:K_5.\end{aligned}\tag{3.34}$$

Avant d'énoncer le théorème nous allons voir que la quantité K_5 peut être simplifiée :

Lemme 27 *Pour une source apériodique, il n'existe pas de couple (β_i,β_j) vérifiant les trois conditions $p^{-\beta_i}+q^{-\beta_i}=1$, $p^{-\beta_j}+q^{-\beta_j}=1$ et $\beta_i+\beta_j=-1$.*

Preuve : Soit un couple (β_i,β_j) vérifiant les trois conditions de l'énoncé, on pose $\beta_i=a$ et ainsi $\beta_j=-1-a$. Nous devons résoudre le système

$$\begin{cases}p^{-a}+q^{-a}=1\\p^{1+a}+q^{1+a}=1.\end{cases}$$

En posant $x = p^a$ et $y = q^a$, on obtient l'équation du second degré $qy^2 - 2qy + 1 = 0$. Cette équation a deux solutions : $y = 1 + i\sqrt{p/q}$ et $y = 1 - i\sqrt{p/q}$ et toutes deux ont pour module $1 + p/q$. Si on écrit $a = s + it$, le module de q^a vaut $q^s = \exp(s \log q)$. Par conséquent, $s \log q = \log(1 + p/q) = \log((q + p)/q) = -\log q$ et $s = -1$. Ce résultat contraint la partie réelle de $\beta_j = -1 - a$ être nulle puisque $\beta_i + \beta_j = -1$. Or nous avons vu dans le lemme 19 de la page 72 qu'il n'existait pas de solution à l'équation $p^{-s} + q^{-s} = 1$ de partie réelle nulle pour une source apériodique ◄

Le lemme précédent permet de simplifier l'écriture de la constante K_5 en

$$K_5 = \frac{2h_2}{h^3} + \frac{2(1+\gamma)}{h^2} - \frac{1}{h}.$$

Nous pouvons ainsi énoncer le théorème sur la variance de la longueur de cheminement dans un trie.

Théorème *Le comportement asymptotique de la variance de la longueur de cheminement d'un trie construit sur des textes engendrés par un modèle de source sans mémoire (p, q) et un modèle de Poisson de paramètre z sur le nombre de textes est*

$$\mathbb{V}_{\mathcal{P}(z)}(L) = \frac{1}{h^2} z \log^2 z + K_5 z \log z + O(z),$$

où

$$K_5 = \frac{2h_2}{h^3} + \frac{2(1+\gamma)}{h^2} - \frac{1}{h}.$$

Le théorème de dépoissonisation analytique de [Szp01] (théorème 10.14, page 481) permet d'obtenir le comportement asymptotique de la longueur de cheminement quand le nombre de chaîne suit un modèle de Bernoulli en connaissant le comportement asymptotique sous un modèle de Poisson de paramètre z. Les hypothèses du théorème sont vérifiées car les fonctions $z \log z$ et $z \log^2 z$ ont toutes deux une croissance de l'ordre adéquat. Nous obtenons alors

Théorème *Le comportement asymptotique de la variance de la longueur de cheminement d'un trie construit sur des textes engendrés par un modèle de source sans mémoire (p, q) et un modèle de Bernoulli de paramètre n sur le nombre de textes est*

$$\mathbb{V}_n(L) = \frac{h_2 - h^2}{h^3} n \log n + O(n).$$

3.4 Série génératrice comptant les co-occurrences de deux motifs

Dans cette section, nous obtenons les séries génératrices des textes contraints par le nombre d'occurrence de deux motifs distincts. Plus spécifiquement nous trouvons les expressions de quatre séries génératrices : celle des textes avec une unique occurrence du motif w et aucune du motif w' (et vice-versa), celle des textes avec une unique occurrence de w et w' et celle des textes sans aucune occurrence ni de w, ni de w'.

Notre méthodologie est celle déjà utilisée dans la section 2.2 du chapitre 2. Elle se base sur la méthode combinatoire introduite par Guibas et Odlyzko [GO81b] qui consiste à décomposer

les textes et à obtenir un système entre différents langages de textes. L'utilisation du dictionnaire entre les langages et les séries formelles permet ensuite d'aboutir aux séries génératrices recherchées.

Dans le chapitre précédent, nous avons exprimé les moyennes de la taille et de la longueur de cheminement d'un arbre des suffixes en fonction des probabilités des textes avec zéro ou une occurrence d'un motif w. Les deux paramètres étaient récrits à l'aide du paramètre $\widehat{N}_w$ qui compte le nombre de suffixes parmi les n premiers suffixes de la chaîne qui commencent par le motif w (soit aussi le nombre d'occurrences du motif w parmi les n premières positions du texte). Les variances de la taille et de la longueur de cheminement dans un arbre des suffixes s'expriment aussi en fonction de $\widehat{N}_w$ et par conséquent en fonction des probabilités des textes avec zéro ou une occurrence de motifs. Néanmoins deux motifs entrent en compte dans l'expression de ces probabilités. Ces probabilités interviennent dans le calcul de l'espérance du carré de la taille et de la longueur de cheminement.

L'auto-corrélation est essentielle dans l'expression des séries génératrices des textes du chapitre 2. On cherchait alors à compter le nombre d'occurrences du motif w dans un texte et il fallait prendre en compte le possible chevauchement de deux occurrences de w. Ici nous devons aussi tenir compte de la possibilité que les deux motifs w et w' dont on cherche à compter les occurrences dans le texte aient des occurrences chevauchantes. On perçoit que si w et w' sont très corrélés, la probabilité d'occurrence va en tenir compte. Le cas le plus éclairant étant de prendre w préfixe de w' : l'apparition de w' nous garantit celle de w, et l'apparition de w en une position donnée signifie qu'il ne manque que le suffixe de w' pour avoir w'. Pour pouvoir quantifier cette corrélation entre deux motifs, nous allons étendre la notion de polynôme d'auto-corrélation à un *polynôme de corrélation* entre deux chaînes. Il nous faut aussi vérifier que si $w = w'$, on retrouve le désormais classique polynôme d'auto-corrélation. La définition du polynôme d'auto-corrélation vient naturellement quand décompose les textes combinatoirement.

L'ensemble des textes sans occurrence ni de w, ni de w' est noté $\mathcal{N}_{0,0} = \mathcal{N}$, sa série génératrice est donnée par le lemme 28 de la page 92. L'ensemble des textes avec une unique occurrence de w et aucune occurrence de w' est noté $\mathcal{N}_{1,0}$, et symétriquement, l'ensemble des textes avec une unique occurrence de w' et aucune occurrence de w est noté $\mathcal{N}_{0,1}$. Les deux séries génératrices sont exprimées dans le lemme 29 de la page 95. Les deux langages $\mathcal{N}_{1,0}$ et $\mathcal{N}_{0,1}$ sont obtenus de manière identique nous n'y consacront qu'une seule section. L'ensemble des textes ayant une unique occurrence de w et de w' est noté $\mathcal{N}_{1,1}$ et sa série génératrice est obtenue au lemme 30 de la page 97. Le cas particulier où w est facteur de w' est traité séparément dans le lemme 31 de la page 98.

Dans la suite de cette section, nous aurons aussi besoin de $\mathcal{T}_w$ (resp. $\mathcal{T}_{w'}$) l'ensemble des textes avec une unique occurrence de w (resp. w'), pour lesquels cette occurrence est finale et qui ne contiennent aucune occurrence de w' (resp. w). L'ensemble $\mathcal{C}_w$ est l'ensemble d'auto-corrélation du motif w, c'est-à-dire l'ensemble des u de taille plus petite que w et tels qu'il existe v vérifiant $w.u = v.w$.

3.4.1 Aucune occurrence ni de w, ni de w'

Dans cette section, nous déterminons la série génératrice comptant les textes sans aucune occurrence ni de w, ni de w'.

Si on concatène le motif w à un texte de l'ensemble $\mathcal{N}$ (c'est-à-dire un texte sans aucune occurrence ni de w, ni de w'), deux configurations distinctes se présentent : soit le premier (en lisant le texte de gauche à droite) mot que l'on rencontre (parmi w et w') dans les textes de $\mathcal{N}.w$ est w, soit c'est w'. Le texte appartient forcément à l'un de ces deux ensembles puisque w

apparait trivialement dans $\mathcal{N}.w$. Pour que les deux ensembles soient disjoints, il est nécessaire que w et w' ne puissent commencer en même temps. On rajoute donc la condition que w n'est pas préfixe de w' et *vice-versa*.

Dans le cas où w' arrive en premier dans le texte de $\mathcal{N}.w$ (devant w), cette occurrence de w' chevauche l'occurrence finale de w (sinon cela contredirait la définition de $\mathcal{N}$). Le texte de $\mathcal{N}.w$ se décompose alors en un texte qui finit par w', ne contient qu'une seule occurrence de w' (occurrence finale) et aucune de w, et un suffixe v de w. Le suffixe v vérifie en plus la propriété : il existe u tel que $w'.v = u.w$. Il semble naturel d'introduire l'ensemble $\mathcal{C}(w', w)$ des textes v (de taille inférieure à w, pour qu'on ait un chevauchement entre w et w') qui vérifient $w'.v = u.w$ pour un certain motif u.

Définition 10 *L'ensemble*

$$\mathcal{C}(w', w) := \{v : |v| < |w| \text{ et il existe } u \text{ tel que } w'.v = u.w\} \tag{3.35}$$

est appelé ensemble de corrélation *entre les motifs w' et w.*

Remarques : Quand $w = w'$, on retrouve la définition de l'ensemble d'auto-corrélation. La notation n'est pas symétrique en w et w'.

Nous pouvons écrire l'équation sur les langages

$$\mathcal{N}.w = \mathcal{T}_w.\mathcal{C}_w + \mathcal{T}_{w'}.\mathcal{C}(w', w)$$

et symétriquement, il vient

$$\mathcal{N}.w' = \mathcal{T}_{w'}.\mathcal{C}'_w + \mathcal{T}_w.\mathcal{C}(w, w').$$

Si on rajoute une lettre à un texte de $\mathcal{N}$, soit on ne crée aucune occurrence ni de w, ni de w', soit on crée une occurrence finale de w (et aucune de w'), soit une occurrence finale de w' (et aucune de w). L'inclusion réciproque est facile.

À partir de ces trois équations, on obtient le système sur les langages suivant :

$$\begin{cases} \mathcal{N}.\mathcal{A} + \epsilon = \mathcal{N} + \mathcal{T}_w + \mathcal{T}_{w'} \\ \mathcal{N}.w = \mathcal{T}_w.\mathcal{C}_w + \mathcal{T}_{w'}.\mathcal{C}(w', w) \\ \mathcal{N}.w' = \mathcal{T}_{w'}.\mathcal{C}_{w'} + \mathcal{T}_w.\mathcal{C}(w, w'). \end{cases} \tag{3.36}$$

Le dictionnaire entre ensembles de langages et séries génératrices permet d'obtenir un système d'équations sur les séries génératrices :

$$\begin{cases} N(z)z + 1 = N(z) + T_w(z) + T_{w'}(z) \\ N(z)p_w z^{|w|} = T_w(z)c[w](z) + T_{w'}(z)c[w', w](z) \\ N(z)p_{w'} z^{|w'|} = T_{w'}(z)c[w'](z) + T_w(z)c[w, w'](z), \end{cases}$$

où les séries génératrices $N(z)$, $T_w(z)$ et $T_{w'}(z)$ sont relatives aux ensembles de textes $\mathcal{N}$, $\mathcal{T}_w$ et $\mathcal{T}_{w'}$ et $c[w', w](z)$ (noté plus simplement $c[w', w]$) est la série génératrice de l'ensemble de corrélation. Il n'y a qu'un nombre fini de textes v possibles et donc $c[w', w]$ est un polynôme.

Définition 11 *Le polynôme $c[w', w](z)$ est le* polynôme (probabilisé) de corrélation *de w' avec w.*

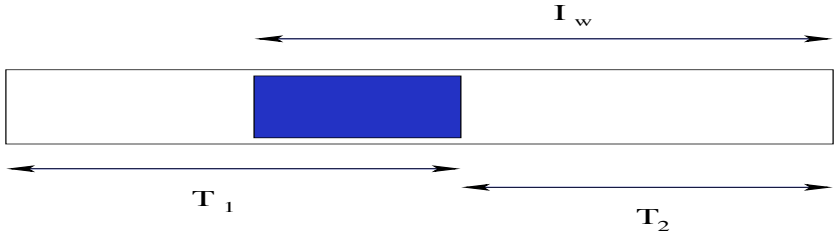


FIG. 3.3 – Découpage des textes avec une unique occurrence de w et aucune de w'

Nous résolvons le système sur les séries génératrices avec les outils classiques. On pose $k = |w|$ et $l = |w'|$. Nous obtenons la relation

$$T_w(z) = T_{w'}(z)\frac{z^k p_w c[w'] - z^l p_{w'} c[w', w]}{z^l p_{w'} c[w] - z^k p_w c[w, w']}.$$

et en posant

$$D[w, w'](z) := c[w'](z)c[w](z) - c[w', w](z)c[w, w'](z). \tag{3.37}$$

Lemme 28 *La série génératrice comptant les textes sans occurrence, ni de w, ni de w' et où z marque la taille du texte vaut*

$$N[w, w'](z) = \frac{D[w, w'](z)}{(1-z)D[w, w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_{w'} z^l c[w', w] - p_w z^k c[w, w'])}. \tag{3.38}$$

La série génératrice $T_w(z)$ (resp, $T_{w'}(z)$) avec une occurrence finale et unique de w (resp. w') et aucune occurrence de w (resp. w) vaut

$$T_w(z) = \frac{z^k p_w c[w'] - z^l p_{w'} c[w', w]}{(1-z)D[w, w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_{w'} z^l c[w', w] - p_w z^k c[w, w'])},$$

et symétriquement

$$T_{w'}(z) = \frac{z^l p_{w'} c[w] - z^k p_w c[w, w']}{(1-z)D[w, w'] + (p_{w'} z^l c[w] + p_w z^k c[w'] - p_w z^k c[w, w'] - p_{w'} z^l c[w', w])}.$$

Remarques : L'expression est symétrique entre w et w' ce qui est logique puisque l'expression de l'ensemble $\mathcal{N}$ ne fait pas de distinction entre w et w'. Les expressions des trois séries génératrices possèdent le même dénominateur.

3.4.2 Une seule occurrence de w et aucune de w'

Dans cette section on cherche l'expression de la série génératrice $N_{1,0}(z)$ de l'ensemble $\mathcal{N}_{1,0}$ des textes qui ne contiennent qu'une seule occurrence de w et dans lesquels w' n'apparaît pas.

Soit $\mathfrak{T}$ un texte de $\mathcal{N}_{1,0}$, il se décompose (cf. figure 3.3) en deux parties : la partie $\mathfrak{T}_1$ qui va du début du texte à la fin de l'unique occurrence de w et le reste $\mathfrak{T}_2$ du texte. Le texte $\mathfrak{T}_2$ ne peut s'identifier à un texte de l'ensemble $\mathcal{N}$ (textes qui ne contiennent aucune occurrence ni de w, ni de w'). Si, effectivement, il n'y a aucune occurrence de w ou de w' dans le texte $\mathfrak{T}_2$, il y a, en plus, la condition que $w.\mathfrak{T}_2$ ne contient qu'une unique occurrence de w et aucune de w'. Plus précisément, cela signifie que la concaténation de w avec $\mathfrak{T}_2$ ne peut pas créer d'occurrence de w supplémentaire, ni d'occurrence de w' dans le texte $\mathfrak{T}$. Cela se traduit par « $w.\mathfrak{T}_2$ est un

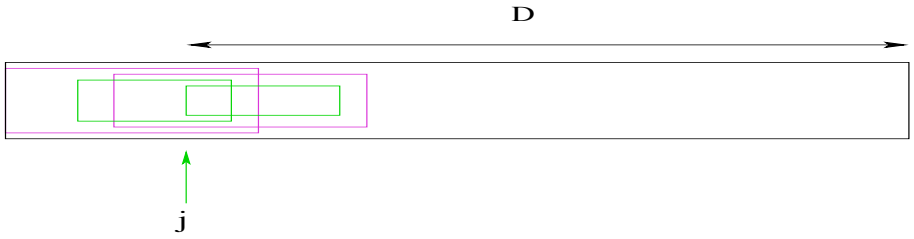


FIG. 3.4 – Apparitions des motifs w et w'

texte avec une unique occurrence de w, cette occurrence est initiale et w' n'apparaît pas dans le texte. » On connait déjà (presque) la série génératrice de ces textes. La série génératrice des textes du type $\mathfrak{T}_1$ (textes avec une seule occurrence de w, occurrence finale et aucune occurrence de w') est aussi connue puisque c'est la série génératrice $T_w(z)$.

Nous déterminons la série génératrice des textes qui ont une occurrence initiale et unique de w et aucune occurrence de w'. On note $\mathcal{I}_w$ l'ensemble des textes qui vérifient cette condition. On cherche un système de langages qui fasse intervenir l'ensemble $\mathcal{I}_w$. Si on rajoute une lettre de l'alphabet $\mathcal{A}$ devant un texte de $\mathcal{N}$, alors on a

- soit un texte qui commence par w, qui ne contient aucune autre occurrence de w, et aucune occurrence de w' (sinon w' serait préfixe de w ou le contraire ce qui est explicitement interdit),
- soit un texte qui a une unique et initiale occurrence de w' et aucune de w',
- soit un texte sans aucune occurrence ni de w, ni de w'.

Ces trois ensembles sont disjoints puisque w ne peut être préfixe de w', ni w' préfixe de w. On note aussi que le mot vide ϵ est toujours un élément de $\mathcal{N}$. Nous avons obtenu une inclusion, mais la réciproque est facile et on aboutit à l'équation

$$\mathcal{A}.\mathcal{N} + \epsilon = \mathcal{N} + \mathcal{I}_w + \mathcal{I}_{w'}. \tag{3.39}$$

D'autre part si on rajoute le motif w devant un texte de l'ensemble $\mathcal{N}$, il existe une dernière position j (figure 3.4) parmi les $|w|$ premières qui voit commencer une occurrence du motif w ou du motif w' (on sait qu'il en existe au moins une, quitte à ce que ce soit la première position du texte). Le texte de $w.\mathcal{N}$ se découpe en une partie $\mathfrak{D}$ qui commence à partir de la position j et va jusqu'à la fin du texte et une partie u qui est un préfixe de w.

Soit le texte $\mathfrak{D}$ commence (à la position j) par le motif w et c'est alors un texte de $\mathcal{I}_w$ (occurrence de w unique et initiale, pas de w') et le texte u est un préfixe d'auto-corrélation de w, c'est-à-dire un motif u (de taille inférieure à $|w|$) pour lequel il existe v qui vérifie $u.w = w.v$. On note $\check{\mathcal{C}}_w$ l'ensemble des u vérifiant la condition précédente.

Soit le texte $\mathfrak{D}$ commence par le motif w', c'est alors un texte de l'ensemble $\mathcal{I}_{w'}$, le préfixe u du motif w vérifie la condition $u.w' = w.v$ où v est un suffixe de w' comme on peut le voir dans la figure 3.5. On note $\check{\mathcal{C}}(w, w')$ l'ensemble des u qui vérifient cette condition.

La réciproque est facile et on écrit l'équation

$$w.\mathcal{N} = \check{\mathcal{C}}_w \mathcal{I}_w + \check{\mathcal{C}}(w, w')\mathcal{I}_{w'}, \tag{3.40}$$

on a l'équation analogue avec w' :

$$w'.\mathcal{N} = \check{\mathcal{C}}_{w'} \mathcal{I}_{w'} + \check{\mathcal{C}}(w', w)\mathcal{I}_w. \tag{3.41}$$

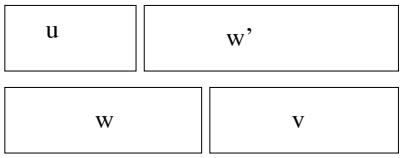


FIG. 3.5 – Pour $\mathcal{C}(w, w')$ on cherche les v qui vérifient cette configuration et pour $\check{\mathcal{C}}(w, w')$, les u

On obtient donc un système de trois équations sur les langages très similaire à celui qui nous a servi à déterminer la série génératrice $N(z)$. Avant de pouvoir résoudre ce système de trois équations à trois inconnues, exprimons les séries génératrices de probabilités des ensembles $\check{\mathcal{C}}_w$ et $\check{\mathcal{C}}(w, w')$ (on trouvera de manière analogue celles pour $\check{\mathcal{C}}_{w'}$ et $\check{\mathcal{C}}(w', w)$).

À chaque u de $\check{\mathcal{C}}_w$ correspond un unique v de même taille qui est un élément de l'ensemble d'auto-corrélation du motif w. De plus, on s'est placé dans un modèle de source sans mémoire, donc la probabilité du motif v est la même que celle du motif u ($p_{u.w} = p_u p_w = p_{w.v} = p_w p_v$). Ainsi la série génératrice $\check{c}[w](z)$ de l'ensemble des motif u (de taille inférieure à $|w|$) pour lesquels il existe un v vérifiant $u.w = w.v$ est le polynôme d'auto-corrélation $c[w](z)$ tel qu'il a été défini dans la chapitre précédent.

Comme dans le paragraphe précédent, on remarque qu'il y a bijection entre les éléments de $\mathcal{C}(w, w')$ et ceux de $\check{\mathcal{C}}(w, w')$: s'il existe un v de taille inférieure à $|w'|$ qui vérifie $u.w' = w.v$, c'est-à-dire un motif de l'ensemble $\mathcal{C}(w, w')$, alors il existe u qui vérifie la condition $u.w' = w.u$ et de taille inférieure à $|w|$. Cela se traduit au niveau analytique dans la série génératrice $\check{c}[w, w'](z)$ de l'ensemble $\check{\mathcal{C}}(w, w')$ (qui est en fait un polynôme). S'il y a un monôme $p_v z^{|v|}$ dans le polynôme de corrélation $c[w, w'](z)$, alors il existe un monôme $p_u z^{|u|}$ dans le polynôme de corrélation $\check{c}[w, w'](z)$ avec les relations $p_u = p_u p_w / p_{w'}$ et $|u| = |v| + k - l$. Ainsi,

$$\check{c}[w, w'](z) = \frac{p_w z^{k-l}}{p_{w'}} c[w, w'](z). \tag{3.42}$$

On regroupe toutes ces informations dans le système

$$\begin{cases} zN(z) + 1 = N(z) + I_w(z) + I_{w'}(z) \\ p_w z^k N(z) = c[w](z) I_w(z) + \frac{p_w z^{k-l}}{p_{w'}} c[w, w'](z) I_{w'}(z) \\ p_{w'} z^l N(z) = c[w'](z) I_{w'}(z) + \frac{p_{w'} z^{l-k}}{p_w} c[w', w](z) I_w(z), \end{cases} \tag{3.43}$$

où $I_w(z)$ (resp. $I_{w'}(z)$) est la série génératrice des textes avec une occurrence de w (resp. w') unique et initiale et aucune occurrence de w' (resp. w).

On a une relation

$$I_w(z) = I_{w'} \frac{p_w z^k (c[w'] - c[w, w'])}{p_{w'} z^l (c[w] - c[w', w])}$$

entre les deux séries génératrices I_w et $I_{w'}$. Les calculs sont fastidieux et nous permettent d'obtenir

$$I_w(z) = \frac{p_w z^k (c[w'] - c[w, w'])}{(1 - z) D[w, w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_w z^k c[w, w'] - p_{w'} z^l c[w', w])}, \tag{3.44}$$

et par symétrie

$$I_{w'}(z) = \frac{p_{w'} z^l (c[w] - c[w', w])}{(1 - z) D[w, w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_w z^k c[w, w'] - p_{w'} z^l c[w', w])}. \tag{3.45}$$

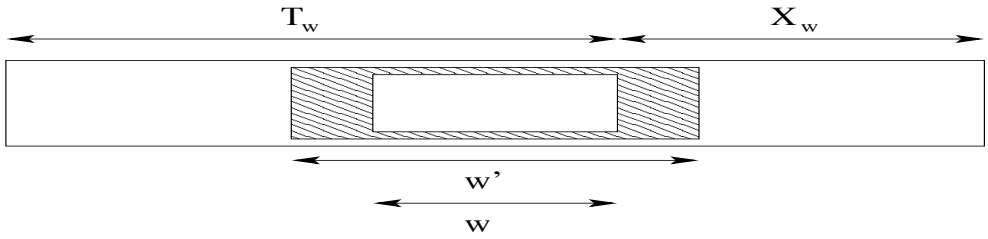


FIG. 3.6 – Cas problématique quand w est facteur de w'

Je rappelle que l'objectif de cette section est d'obtenir la série génératrice des textes avec une unique occurrence de w et aucune de w'. Les textes de l'ensemble $\mathcal{N}_{1,0}$ se décomposent en un texte de $\mathcal{T}_w$ et un texte d'un ensemble $\mathcal{X}_w$ vérifiant $w.\mathcal{X}_w = \mathcal{I}_w$. Néanmoins il faut faire attention : la réciproque n'est pas vraie. Si le motif w' a w comme facteur, on peut avoir w' qui apparaît dans les textes de $\mathcal{T}_w.\mathcal{X}_w$ sans jamais apparaître ni dans $\mathcal{T}_w$, ni dans $\mathcal{X}_w$ comme le montre la figure 3.6. Pour pallier ce problème, on interdit à w d'être un facteur de w' et réciproquement. On traitera ce cas à part. Le cas où, par exemple, w est facteur de w' englobe le cas où w est préfixe de w' que nous avions déjà exclu pour obtenir des équations sur les langages non-ambigus. Une fois ce cas gênant mis de côté, on a bien la relation

$$\mathcal{N}_{1,0} = \mathcal{T}_w.\mathcal{X}_w,$$

et ainsi

Lemme 29 *Soit w et w' deux motifs tels que w ne soit pas facteur de w'. La série génératrice des textes dont la taille est comptée par la variable z avec une unique occurrence de w et aucune de w' s'écrit*

$$N_{1,0}(z) = \frac{(z^k p_w c[w'] - z^l p_{w'} c[w', w])(c[w'] - c[w, w'])}{((1-z)D[w, w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_{w'} z^l c[w', w] - p_w z^k c[w, w']))^2}. \quad (3.46)$$

3.4.3 Une unique occurrence de w et de w'

Dans cette section, nous déterminons la série génératrice $N_{1,1}(z)$ des textes avec une unique occurrence de w et de w'. On suppose que le motif w apparaît en premier dans le texte. On n'oubliera pas de compter le cas symétrique (w' arrivant avant w) pour déterminer $N_{1,1}(z)$. On traite le cas général c'est-à-dire celui où w et w' ne sont pas facteurs l'un de l'autre. L'ensemble des textes avec une occurrence de w et de w' où de plus w apparaît avant w est noté $\mathcal{N}'_{1,1}$.

Un texte de $\mathcal{N}'_{1,1}$ se décompose (cf. figure 3.7) en trois morceaux de texte : $\mathfrak{A}$ un texte de l'ensemble $\mathcal{T}_w$, $\mathfrak{B}$ un texte de $\mathcal{U}_{w,w'}$ et $\mathfrak{C}$ un texte de $\mathcal{I}_{w'}$. L'ensemble $\mathcal{U}_{w,w'}$ est défini comme l'ensemble des textes avec une occurrence de w qui soit unique et initiale et une occurrence de w' unique et finale. Nous ne rajoutons aucune condition sur le chevauchement des motifs w et w' (les deux occurrences peuvent être chevauchantes ou non [comme dans la figure]), mais ni w', ni w ne peuvent être facteur de l'autre.

Les ensembles $\mathcal{I}_{w'}$ et $\mathcal{T}_w$ sont déjà connus ainsi que leurs séries génératrices, mais on ne sait rien (pour l'instant) de $\mathcal{U}_{w,w'}$. Regardons ce qui se passse si on rajoute une lettre à l'extrémité droite de I_w :

– soit l'ajout de cette lettre ne crée aucune occurrence ni de w, ni de w' en suffixe du nouveau texte. On retrouve donc un texte de $\mathcal{I}_w$.

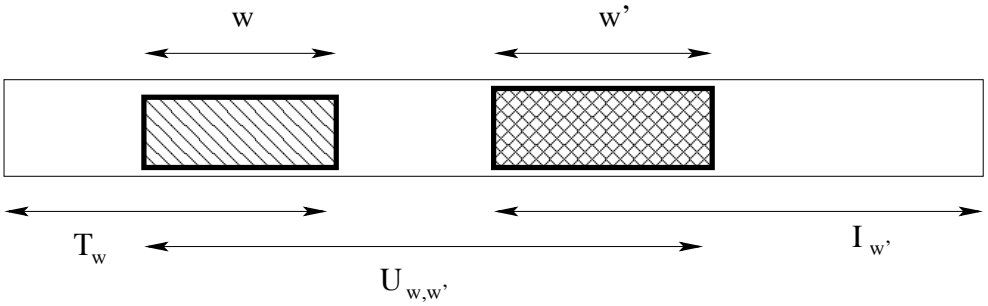


FIG. 3.7 – Texte avec une unique occurrence de w et de w'

– soit on crée une occurrence finale de w' et on a un élément de $\mathcal{U}_{w,w'}$,
– soit on crée une occurrence finale de w et on a un élément de $\mathcal{M}_w$, l'ensemble des textes avec exactement deux occurrences de w : une en suffixe et une en préfixe, et aucune occurrence de w'.

Les trois cas sont disjoints. Pour l'inclusion réciproque, il faut rajouter dans l'équation le motif w qui est un élément de $\mathcal{I}_w$ mais ne peut s'écrire sous la forme $\mathcal{I}_w.\mathcal{A}$. Nous aboutissons à

$$\mathcal{I}_w.\mathcal{A} + w = \mathcal{I}_w + \mathcal{U}_{w,w'} + \mathcal{M}_w. \tag{3.47}$$

On a deux inconnues dans l'équation précédente, il nous faut donc obtenir une seconde équation.

Si on ajoute le motif w' à la fin d'un texte de $\mathcal{I}_w$, deux configurations existent selon que l'on rencontre une deuxième occurrence (éventuelle) de w avant la première occurrence de w' ou non. Soit une deuxième occurrence de w apparaît avant la première occurrence de w' et alors on décompose le texte en un élément de $\mathcal{M}_w$ et un élément de $\mathcal{C}_{w,w'}$, soit la première occurrence de w' arrive avant une (éventuelle) deuxième occurrence de w et le texte de $\mathcal{I}_w.w'$ s'écrit comme le produit d'un texte de $\mathcal{U}_{w,w'}$ et d'un texte de $\mathcal{C}_{w'}$. En fait dans le second cas, cela signifie que w ne réapparaît pas. On rassemble ce découpage dans l'équation

$$\mathcal{I}_w.w' = \mathcal{M}_w.\mathcal{C}_{w,w'} + \mathcal{U}_{w,w'}.\mathcal{C}_{w'}. \tag{3.48}$$

Les équations (3.47) et (3.48) forment un système sur les langages, on passe à un système sur les séries génératrices de probabilités :

$$\begin{cases} zI_w(z) = I_w(z) + U_{w,w'}(z) + M_w(z) \\ I_w(z)p_{w'}z^l = M_w(z)c[w,w'](z) + U_{w,w'}(z)c[w'](z). \end{cases}$$

La résolution du système ne présente pas de problème :

$$\begin{aligned} U_{w,w'}(z) &= I_w(z)\frac{(1-z)c[w,w'] + p_{w'}z^l}{c[w'] - c[w,w']} \\ &= \frac{p_w z^k((1-z)c[w,w'] + p_{w'}z^l)}{(1-z)D[w,w'] + (p_w z^k c[w'] + p_{w'}z^l c[w] - p_w z^k c[w,w'] - p_{w'}z^l c[w',w])}. \end{aligned}$$

Remarque 1 : On aurait pu remplacer la seconde équation (3.48) du système par $\mathcal{I}_w.w = \mathcal{U}_{w,w'}\mathcal{C}_{w',w} + \mathcal{M}_w\mathcal{C}_w$ et aboutir (heureusement) au même résultat pour la série génératrice $U_{w,w'}(z)$.

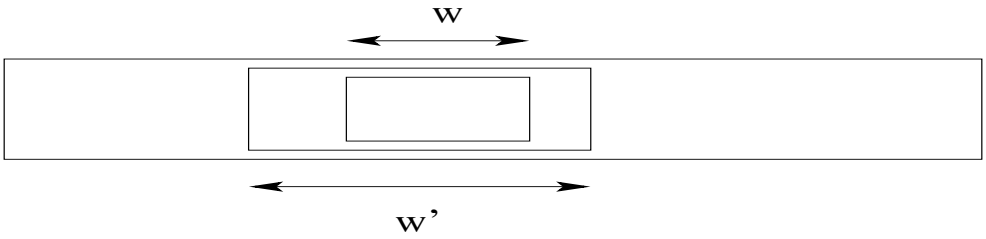


FIG. 3.8 – w est facteur du motif w'

Remarque 2 : La série génératrice $M_w(z)$ a été éliminée de notre système, on aurait toutefois pu se servir de son expression : elle est connue depuis les résultats de Régnier et Szpankowski dans [RS97, RS98].

Montrons que les textes de $w.w'.\mathcal{N}'_{1,1}$ sont en bijection avec les textes de $\mathcal{T}_w.\mathcal{U}_{w,w'}.\mathcal{I}_{w'}$. Si un texte appartient à $\mathcal{T}_w.\mathcal{U}_{w,w'}.\mathcal{I}_{w'}$, on peut en couper les extrémités w et w' de $\mathcal{U}_{w,w'}$ pour obtenir un texte de $\mathcal{N}'_{1,1}$. On aura une unique occurrence de w et de w' et il ne peut y avoir d'autres occurrences sans contrevenir à la condition w n'est pas facteur de w' et *vice-versa*. Dans le cas de émis par une source sans mémoire, cette bijection permet d'affirmer d'obtenir la série génératrice des probabilités des textes de $\mathcal{N}'_{1,1}$. Tu te rappelleras, ô lecteur, que je n'ai regardé ici que les textes avec une unique occurrence de w et de w' pour lesquels w arrive avant w'. Ainsi la série génératrice du « vrai » ensemble $\mathcal{N}_{1,1}$ est définie par l'équation

$$p_w z^k p_{w'} z^l N_{1,1}(z) = T_w(z) U_{w,w'}(z) I_{w'}(z) + T_{w'}(z) U_{w',w}(z).I_w(z). \tag{3.49}$$

Nous utilisons notre connaissance des séries génératrices $T_w(z)$ et $U_{w,w'}(z)$ pour obtenir le lemme

Lemme 30 *La série génératrice qui compte la probabilité des textes de taille n avec une unique occurrence de w et une unique occurrence de w' vaut*

$$N_{1,1}(z) = \frac{(p_w z^k c[w'] - z^l p_{w'} c[w', w])(c[w] - c[w', w])((1-z)c[w, w'] + p_{w'} z^l)}{((1-z)D[w, w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_w z^k c[w, w'] - p_{w'} z^l c[w', w]))^3}. \tag{3.50}$$

3.4.4 Cas particulier d'un motif facteur d'un autre

Nous traitons maintenant du cas particulier qui a été mis de côté dans l'étude générale : si w est un facteur de w'. Le cas où w' est un facteur de w est symétrique. Ce cas a été mis à part car la décomposition présentée dans le cas général devient ambiguë. Nous cherchons, comme dans le cas général, à determiner les séries génératrices de probabilités des textes où il n'y a aucune occurrence ni du motif w, ni du motif w', où il y a une occurrence de w et aucune de w', une occurrence de w' et aucune de w et enfin où il y a une unique occurrence de w et de w'. Les relations particulières entre les deux motifs vont être abondamment exploitées. Nous ne regardons que le cas où w apparaît en facteur de w' à une seule reprise (cf. figure 3.8). Dans les cas où il y a plusieurs apparitions de w dans w', les résultats sont soit identiques, soit trivaux et nous y consacreront un bref paragraphe. Le cas où w est facteur de w' inclus le cas où w est préfixe de w'.

Si w est un facteur de w', dire qu'un texte ne contient aucune occurrence ni de w, ni de w' est équivalent à dire que le texte ne contient aucune occurrence de w. La série génératrice $N(z)$

est donc celle déterminée à l'équation (2.6) de la page 41 :

$$\mathcal{N}_{1,1}(z) = N(z) = \frac{c[w](z)}{p_w z^{|w|} + (1-z)c[w](z)}.$$

De manière triviale la série génératrice des textes avec une unique occurrence de w' et aucune de w est identiquement nulle puisqu'on ne peut avoir w' sans avoir w.

Déterminons la série génératrice des textes avec des occurrences de w et de w' uniques : dans le cas où w est facteur de w' et où w apparaît à une seule reprise, le motif w' ne peut apparaître qu'au plus une fois (sinon il entraînerait l'apparition de w à plus d'une reprise). Donc l'ensemble des textes avec une occurrence de w (noté antérieurement par $\mathcal{O}_w$) se décompose de manière non-ambiguë en l'ensemble des textes avec une unique occurrence de w et une unique occurrence de w' (noté $\mathcal{N}^f_{1,1}$, où l'indice f signifie facteur) et l'ensemble des textes avec une seule apparition de w et aucune de w' (noté $\mathcal{N}^f_{1,0}$), soit

$$\mathcal{O}_w = \mathcal{N}^f_{1,1} + \mathcal{N}^f_{1,0}. \tag{3.51}$$

L'ensemble des textes avec une seule apparition de w et sa série génératrice ont été préalablement déterminés. En passant aux séries génératrices, on obtient

$$O_w(z) = \frac{p_w z^k}{(z^k p_w + c[w](z)(1-z))^2} = N^f_{1,1}(z) + N^f_{1,0}(z).$$

Il nous faut maintenant trouver l'une des deux séries génératrices de probabilités manquantes. On s'occupe de l'ensemble $\mathcal{N}^f_{1,1}$. S'il y a une occurrence de w' alors nécessairement il y en a une de w, l'ensemble $\mathcal{N}^f_{1,1}$ se définit aussi comme celui des textes avec une unique occurrence de w' et aucune occurrence de w (en dehors de l'occurrence triviale conséquente à l'apparition de w'). L'idée est de masquer lors de la mise en équation l'occurrence de w liée à l'apparition de w'.

On reprend le résultat symétrique du lemme 29 provenant de la résolution du système (3.43) et on aboutit à

$$N^f_{1,1}(z) = N_{0,1}(z) = \frac{(z^l p_{w'} c[w] - z^k p_w c[w,w'])(c[w] - c[w',w])}{((1-z)D[w,w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_{w'} z^l c[w',w] - p_w z^k c[w,w']))^2}.$$

Lemme 31 *Dans le cas où w est un facteur de w', les séries génératrices des textes sont*

$$N_{0,0}(z) = \frac{c[w](z)}{p_w z^{|w|} + (1-z)c[w](z)},\ N_{0,1}(z) = 0$$

$$N_{1,1}(z) = \frac{(z^l p_{w'} c[w] - z^k p_w c[w,w'])(c[w] - c[w',w])}{((1-z)D[w,w'] + (p_w z^k c[w'] + p_{w'} z^l c[w] - p_{w'} z^l c[w',w] - p_w z^k c[w,w']))^2},$$

$$\text{et } N_{1,0}(z) = \frac{p_w z^k}{(z^k p_w + c[w](z)(1-z))^2} - N_{1,1}(z).$$

Expliquons rapidement ce qui va se passer quand le motif w apparaît à plusieurs reprises dans le texte w', mettons deux fois. La série génératrice des textes avec aucune occurrence ni de w, ni de w' est identique au cas où w apparaît en facteur de w' à une seule reprise. Pour les textes avec 1 occurrence de w et aucune de w', cela correspond aux textes avec une seule occurrence de w puisqu'il faut au moins deux occurrences de w pour que w' apparaisse. Donc on écrit $\mathcal{O}_w = \mathcal{N}_{1,0}$, avant de passer aux séries génératrices. Il n'est toujours pas possible d'avoir des textes avec une occurrence de w' et aucune occurrence de w. La nouveauté vient des textes avec une unique occurrence de w et de w' : il n'y en a pas !

3.5 Conclusion

Dans ce chapitre nous avons montré que la variance de la taille S d'un trie construit sur un nombre de textes suivant une loi de Poisson de paramètre z et engendrés par une source apériodique sans mémoire (p, q) se comporte asymptotiquement en $O(z)$:

$$\mathbb{V}_{\mathcal{P}(z)}(S) = O(z). \tag{3.52}$$

Nous avons aussi montré que pour la longueur de cheminement

Théorème *Le comportement asymptotique de la variance de la longueur de cheminement L d'un trie construit sur des textes engendrés par un modèle de source apériodique sans mémoire (p, q) et un modèle de Poisson de paramètre z sur le nombre de textes est*

$$\mathbb{V}_{\mathcal{P}(z)}(L) = \frac{1}{h^2} z \log^2 z + K_5 z \log z + O(z),$$

et sous un modèle de Bernoulli de paramètre n sur le nombre de textes

$$\mathbb{V}_n(L) = K_6 n \log n + O(n),$$

où

$$K_5 = \frac{2h_2}{h^3} + \frac{2(1+\gamma)}{h^2} - \frac{1}{h}, \quad K_6 = \frac{h_2 - h^2}{h^3} \quad \text{et } h_2 := p \log^2 p + q \log^2 q.$$

Les séries génératrices comptant les textes selon leur taille et une ou zéro occurrence de deux motifs distincts ont aussi été obtenues. Ce chapitre est la première étape d'un travail devant permettre de déterminer le comportement asymptotique de la variance de la taille et de la longueur de cheminement dans un arbre des suffixes.

Nous devons d'abord déterminer le comportement asymptotique complet de la variance de la taille et de la longueur de cheminement dans un trie dans le modèle sans mémoire. Pour se libérer de calculs pesants, il doit être possible de recourir à des méthodes automatiques. D'autre part, le coefficient du terme linéaire et les termes sous-linéaires du comportement asymptotique de la variance de la taille et de la longueur de cheminement n'ont pas été calculés. Les résultats antérieurs de [KP91] pour la taille et [KPS89] pour la longueur de cheminement) sont valables pour une source sans mémoire symétrique. Nos résultats sont valables pour une source sans mémoire biaisée. Les résultats [JR87, RJ89] sont énoncés pour la taille d'un trie sous un modèle de source biaisée. Le résultat de [JR88] sur la variance de la longueur de cheminement dans un trie est énoncé sous un modèle de Poisson et pour une source sans mémoire quelconque. L'analyse proposée dans ce chapitre rectifie le coefficient d'ordre $z \log z$ et offre en plus le comportement asymptotique sous modèle de Bernoulli.

Pour les deux paramètres, notre méthode semble s'adapter, au moins pour la détermination des formules des différentes sommes au cas de sources de Markov. Une étude dans le cadre plus général des sources dynamiques de Vallée [Val01] peut aussi être envisagée. Les comportements asymptotiques de la variance de la taille et de la longueur de cheminement d'un trie permettent d'affirmer que ces deux paramètres sont concentrés, c'est-à-dire qu'ils sont proches de leur moyenne.

La méthode proposée dans le chapitre précédent a donné de bons résultats. Elle consistait à d'abord déterminer certaines séries génératrices puis à en extraire le comportement asymptotique des coefficients à l'aide de la connaissance des pôles dominants et d'outils d'analyse complexe.

Cela nous permettait d'exprimer l'espérance de la taille et de la longueur de cheminement d'un arbre des suffixes. Les séries génératrices que nous venons d'obtenir dans la section précédente sont particulièrement complexes et la localisation de leurs pôles dominants est ardue. Ce travail reste à faire pour permettre d'aboutir à une expression de la variance de la taille et de la longueur de cheminement dans un arbre des suffixes. La dernière étape est de comparer l'expression de la variance pour la taille et la longueur de cheminement dans un arbre des suffixes et dans un trie. Les résultats obtenus sont déjà un progrès dans la détermination du comportement asymptotique de la variance de la taille et de la longueur de cheminement d'un arbre des suffixes.

Un travail a été fait sur la généralisation de la section 3.4. La série génératrice trivariée qui compte les textes par leur taille, le nombre d'occurrences d'un motif w et le nombre d'occurrences d'un autre motif w' ainsi que la série génératrice multivariée comptant les occurrences de plusieurs motifs ont été obtenues par ailleurs. Ces résultats ne sont pas inclus dans la thèse.

Chapitre 4

Anti-dictionnaire

La compression sans perte par anti-dictionnaire, couramment appelée DCA, est une technique originale et récente introduite à la fin des années quatre-vingt dix. Son principe est de construire le dictionnaire de certains mots n'apparaissant pas dans le texte. Dans notre étude, les textes sont engendrés par un modèle sans mémoire. Nous montrons qu'asymptotiquement le nombre moyen de mots dans l'anti-dictionnaire sur l'ensemble des textes de taille n se comporte en $Kn/h + o(n)$ où la constante K est determinée explicitement et h est l'entropie du modèle probabiliste. Nous utilisons un découpage des motifs selon leur longueur (longs, courts et intermédiaires) déjà vu dans le chapitre 3.

The lossless data compression scheme using anti-dictionaries called DCA is quite novel and dates back to the late nineties. Its principle is to build the dictionary of a certain set of words that do **not** occur in the text. In our study the model for generating the texts is memoryless. We show in this chapter that asymptotically the average over all texts of size n of the number of words in the anti-dictionary is $Kn/h + o(n)$ where h is the entropy of the model and K a constant explicity computed. Like in chapter 3 we split the study according to the size of the patterns (long, short and intermediate).

Sommaire

4.1 Introduction

Le principe de l'algorithme de compression de données sans perte LZ'77 (déjà vu dans la section 1.3) consiste à créer un **dictionnaire** des mots déjà rencontrés lors du parcours du texte. Lorsqu'un mot déjà inséré dans le dictionnaire se retrouve lors de la lecture du texte, il est codé par son numéro de référence dans le dictionnaire. Pourvu que le dictionnaire soit implanté avec une structure de données satisfaisante, le numéro de référence est un petit entier. Cette idée de dictionnaire permet de compresser un texte à un niveau proche de l'entropie (c'est-à-dire proche du taux optimal de compression). Lempel et Ziv utilisent dans leur algorithme LZ'77 un arbre

des suffixes comme structure de données pour leur dictionnaire, et dans leur algorithme LZ'78, un arbre digital de recherche pour coder le dictionnaire.

Dans [CMRS99] et [CMRS00], Crochemore, Mignosi, Restivo et Salemi introduisent un nouveau paradigme d'algorithmes de compression de données sans perte appelé `DCA` pour *Data Compression using Antidictionaries*. Cet algorithme se base sur la construction d'un **anti-dictionnaire**. Au lieu de garder en mémoire les mots déjà vus, l'anti-dictionnaire stocke un ensemble de mots non-rencontrés. Ses performances linéaires de compression et de décompression le rendent très attractif. Crochemore et al. montrent que l'algorithme `DCA` atteint asymptotiquement l'entropie sous un modèle de source équilibrées.

Morita et Ota [OM04] utilisent le principe de l'anti-dictionnaire pour comprimer sans perte le résultat d'un électrocardiogramme (ECG). Ils obtiennent un taux de compression amélioré de 10 % par rapport à un algorithme de type Lempel-Ziv (basé sur un dictionnaire).

Le nombre de mots non-rencontrés dans un texte fini est infini. Pour faire fonctionner le principe de compression par anti-dictionnaire, il faut trouver un sous-ensemble fini des mots non-rencontrés : les mots minimaux interdits (*minimal forbidden words* en anglais, ou ici MMI) d'un texte T. Cette notion de minimalité permet d'avoir un ensemble fini de mots non-rencontrés.

Définition 12 *On dit que le mot w est un mot minimal interdit pour un texte T s'il n'apparaît pas dans le texte et si tous ses facteurs y apparaissent.*

La définition d'un MMI fait intervenir tous les facteurs du mot, néanmoins il est plus pratique de recourir à une définition équivalente : le mot w de taille k est un MMI pour le texte T si ses deux facteurs de taille $k-1$ apparaissent dans le texte et que w n'y apparaît pas.

L'anti-dictionnaire bâti sur le texte T est l'ensemble des MMIs du texte T ; sa taille $\mathcal{S}$ est le nombre de mots qui le composent. La taille d'un anti-dictionnaire est donc finie.

Dans [MO04], Morita et Ota ont montré que la taille de l'anti-dictionnaire est toujours plus petite que la taille du dictionnaire associé au même texte.

Dans ce chapitre, je montre que le comportement asymptotique de l'espérance de la taille $\mathcal{S}$ d'un anti-dictionnaire construit à partir d'un texte T de taille donnée n et généré par une source sans mémoire est linéaire. La constante de linéarité est explicitement déterminée. Les textes sont produits par une source sans mémoire (symétrique ou biaisée) sur un alphabet binaire. Les méthodes sont celles, désormais usuelles, de la combinatoire analytique : séries génératrices, transformation de Mellin, et analyse complexe, mais le cœur du problème se cache dans la définition complexe d'un MMI.

4.1.1 Fonctionnement de l'algorithme `DCA`

Nous décrivons le fonctionnement de l'algorithme de compression utilisant l'anti-dictionnaire. Il existe plusieurs implantations possibles de l'algorithme. On distingue trois étapes principales : la construction de l'anti-dictionnaire, la compression du texte et enfin la décompression. Après l'étape de compression du texte T, le code consiste en l'anti-dictionnaire $\mathcal{AD}$ construit sur T, le texte comprimé $\mathcal{T}$ et la longueur n de T. La décompression reconstruit le texte T à partir du code transmis et de l'anti-dictionnaire.

➜ Construction de l'anti-dictionnaire

La construction de l'anti-dictionnaire du texte T se fait avec notre méthode en temps linéaire (le temps de lire le texte). Anticipant sur l'étape suivante de compression, l'anti-dictionnaire est représenté dans une structure de trie.

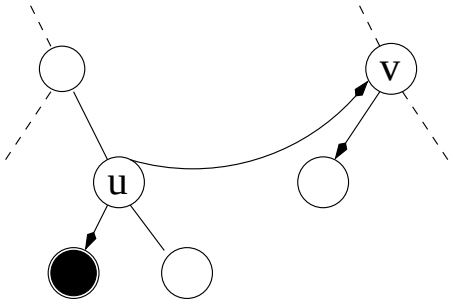


FIG. 4.1 – Le nœud gauche de u est vide et le nœud gauche de v existe donc $u\mathtt{0}$ est un MMI.

Les résultats de Weiner [Wei73], McCreight [McC76] et Ukkonen [Ukk95] montrent que l'on peut construire l'arbre des suffixes d'un texte T de taille n en temps linéaire à l'aide de l'introduction de liens suffixes. La légère modification suivante de l'arbre des suffixes produit sur le texte T permet d'obtenir le trie des MMIs : soit un nœud $u = \alpha v$ sans fils gauche (et/ou droit) avec $\alpha \in \mathcal{A}$. On suit le lien suffixe de u vers v. Si le fils gauche (et/ou droit) de v existe, alors le fils gauche (et/ou droit) de u est un MMI et on le marque comme tel (par une feuille par exemple). Une fois cette phase achevée, toutes les branches de l'arbre qui ne mènent pas à un MMI sont élaguées.

Morita et Ota [MO04] ont explicité l'algorithme « réciproque » `AD2D` permettant de passer du trie formé sur l'anti-dictionnaire d'un texte à l'arbre des suffixes du texte.

➜ Compression

Supposons que nous ayons déjà comprimé le texte $T[0 \cdots i]$. S'il existe un suffixe v du texte $T[0 \cdots i]$ pour lequel $v\mathtt{0}$ (resp. $v\mathtt{1}$) appartient à l'anti-dictionnaire $\mathcal{AD}$ alors la lettre suivante $T[i+1]$ dans le texte est nécessairement `1` (resp. `0`). Cette lettre est déterminée à partir de l'anti-dictionnaire et du texte déjà rencontré, elle est donc redondante. Elle n'est pas notée dans la sortie $\mathcal{T}$ de la compression.

Si par contre il n'existe aucun suffixe v de $T[0 \cdots i]$ tel que $v\mathtt{0}$ ou $v\mathtt{1}$ apparaisse dans l'anti-dictionnaire, on rajoute la lettre $T[i+1]$ dans la sortie $\mathcal{T}$ de l'algorithme de compression.

```
        Compression

T := ε ;
pour i := 0 à n faire
   si pour tout suffixe v de T[0]···T[i], v0 ∉ AD et v1 ∉ AD
      alors T := T.T[i+1] ;
   fin si
fin pour
```

Remarque 1 : Si les mots $v\mathtt{0}$ et $v\mathtt{1}$ appartiennent tous deux à l'anti-dictionnaire alors le texte doit nécessairement se terminer après une occurrence de v puisque ni la lettre `0`, ni la lettre `1` ne peuvent apparaître après le motif v.

Remarque 2 : S'il existe un suffixe v du texte $T[0 \cdots i]$ tel que $w = v\mathtt{0}$ (par exemple) est dans l'anti-dictionnaire alors il n'existe pas d'autre suffixe v' de $T[0 \cdots i]$ pour lequel $v'\mathtt{0}$ (ou $v'\mathtt{1}$) appartient à l'anti-dictionnaire (si v' est plus grand que v par exemple, on sait que à la fois $v\mathtt{0}$ n'appartient pas au texte et qu'il y apparaît en tant que suffixe de $v'\mathtt{0}$).

➜ Décompression

On décrit maintenant le principe de la décompression d'un texte quand on connait la sortie $\mathcal{T}$ de l'algorithme de compression, l'anti-dictionnaire $\mathcal{AD}$ et la longueur n du texte initial. Supposons que l'on ait recréé le texte initial T jusqu'à la position j. S'il existe un suffixe v de ce texte tel que $w = v0$ (resp. $v1$) appartienne à l'anti-dictionnaire alors la lettre $T[j+1]$ est automatiquement déterminée puisque w ne peut apparaître dans le texte T. La lettre $T[j+1]$ est nécessairement `1` (resp. `0`). Il n'est pas besoin de vérifier que le suffixe de taille $k-1$ de w apparaît dans le texte.

La détermination de la lettre $T[j+1]$ à partir de l'anti-dictionnaire n'est possible que parce que l'alphabet est binaire et que l'« autre » lettre est bien définie de manière unique.

Pour tester si le préfixe propre d'un MMI est aussi le suffixe du texte en cours de reconstruction, [CMRS99] utilisent une structure de données issue de la théorie des automates : les transducteurs. Le transducteur qui implante l'anti-dictionnaire est construit en temps linéaire et le test est alors effectué en temps logarithmique. Nous décrivons ici les algorithmes en utilisant les arbres des suffixes avec des liens suffixes. Ces deux structures de données sont très proches et toutes deux adaptées pour gérer l'anti-dictionnaire. Dans les deux cas la décompression du texte se fait en temps linéaire.

```
                Décompression

T := ϵ ; j := 0
tant que |T| < n faire
   si pour tout suffixe v de T, v0 ∉ AD et v1 ∉ AD
      alors T := T.T[j++] ;
   sinon
      si v0 ∈ AD
        T := T.1 ;
      sinon
        T := T.0 ;
      fin si
   fin si
fin tant que
```

4.1.2 Plan et résultat

Les textes sont produits par une source sans mémoire (symétrique ou biaisée) sur un alphabet binaire. Je montre que le comportement asymptotique de l'espérance de la taille $\mathcal{S}$ d'un anti-dictionnaire construit à partir d'un texte T de taille n est linéaire. La constante de linéarité est explicitement déterminée.

Pour un mot w de taille k, les deux facteurs de taille $k-1$ de w sont $w_L := w_1 \cdots w_{k-1}$, le préfixe de taille $k-1$ de w et $w_R := w_2 \cdots w_k$, le suffixe de taille $k-1$ de w. On introduit le motif $u(w) = u := w_2 \cdots w_{k-1}$, $\alpha = u_1$ et $\beta = u_k$. Soient deux lettres quelconques a et b, le mot $v = aub$ est appelé une *extension* de u, et les lettres a et b sont dites *adjacentes* à u.

La probabilité sur l'ensemble des textes de taille n que le mot w soit un MMI est notée $\mathbb{P}_n(w \in \text{ MMI}) = \mathbb{P}(\{T,\ |T| = n \text{ et } w \in \text{MMI}(T)\})$. Pour alléger la notation, l'indice n sera souvent omis. Dans le reste du chapitre, et sauf mention contraire, les textes auront longueur n.

La taille $\mathcal{S}$ de l'anti-dictionnaire bâti sur un texte T est le nombre de mots dans l'anti-dictionnaire, c'est-à-dire le nombre de MMI dans le texte T :

$$\mathcal{S} = \sum_{w \in \mathcal{A}^\star} [\![w \in \text{ MMI}]\!],$$

où $[\![.]\!]$ est le crochet d'Iverson, une notation de la fonction indicatrice déjà introduite page 20. La moyenne de la taille sur tous les textes de taille n vaut

$$\mathbb{E}_n(\mathcal{S}) = \sum_{w \in \mathcal{A}^\star} \mathbb{E}_n([\![w \in \text{ MMI}]\!]) = \sum_{w \in \mathcal{A}^\star} \mathbb{P}_n(w \in \text{ MMI}). \tag{4.1}$$

Le travail effectué dans ce chapitre fait appel aux mêmes méthodes que dans les chapitres précédents : découpage des motifs w dans la somme (4.1) selon leur taille, utilisation de la combinatoire analytique et majoration des valeurs du polynôme d'auto-corrélation.

On traite d'abord les motifs de petite taille (pour lesquels np_w tend vers l'infini) puis les motifs de grande taille (pour lesquels np_w tend vers zéro). Ces deux ensembles de motifs contribuent en $O(n^\beta)$ avec $\beta < 1$ au comportement asymptotique de l'espérance $\mathbb{E}_n(\mathcal{S})$.

Le traitement des motifs de taille intermédiaire est plus délicat. Il requiert le passage par un modèle de calcul simplifié et approché qui suppose deux hypothèses : premièrement que, pour un MMI w, les occurrences de $u(w)$ sont séparées par au moins deux lettres ; deuxièmement que le nombre d'occurrences du motif u dans un texte de taille n suit une loi de Poisson de paramètre np_u. Nous obtenons sous ce modèle et *via* la transformée de Mellin un comportement asymptotique linéaire pour la taille moyenne d'un anti-dictionnaire. Cela confirme les simulations réalisées dans [OM04]. De plus nous explicitons la constante de linéarité ainsi qu'un développement asymptotique de la taille moyenne sous ce modèle.

Dans une troisième partie (sections 4.6 et 4.7), nous montrons que les deux hypothèses du modèle simplifié ne contribuent à la taille moyenne de l'anti-dictionnaire qu'en un terme sous-linéaire. Le résultat de linéarité obtenu sous le modèle approché est donc valable pour la taille moyenne de l'anti-dictionnaire. Cela permet alors d'énoncer le théorème

Théorème 8 *Soit un modèle probabiliste sans mémoire biaisé (p, q) de génération des textes. Le comportement asymptotique de la moyenne $\mathbb{E}_n(\mathcal{S})$ sur les textes de taille n de la taille de l'anti-dictionnaire vaut dans le cas d'une source périodique*

$$\mathbb{E}_n(\mathcal{S}) = K\frac{n}{h} + \frac{n}{h}\epsilon(n) + o(n) \tag{4.2}$$

et dans le cas d'une source apériodique,

$$\mathbb{E}_n(\mathcal{S}) = K\frac{n}{h} + o(n), \tag{4.3}$$

où

$$K := 2h + (1-p^2)\log(1-p^2) + (1-q^2)\log(1-q^2) + 2(1-pq)\log(1-pq) \text{ et}$$

$$\epsilon(n) := \sum_{k \in \mathbb{Z}^*} n^{-s_k - 1}\Gamma(s_k)\Big[\big[p^{-2s_k} - 2p^{-s_k} + (1-q^2)^{-s_k}\big] + \big[q^{-2s_k} - 2q^{-s_k} + (1-p^2)^{-s_k}\big] + 2\big[(pq)^{-s_k} - q^{-2s_k} - p^{-2s_k} + (1-pq)^{-s_k}\big]\Big].$$

Remarque : Le terme d'anti-dictionnaire est aussi utilisé dans la communauté s'occupant d'indexation de contenu et d'analyse linguistique pour désigner exactement le contraire de ce que nous appelons anti-dictionnaire : les mots apparaissant de manière trop fréquente pour avoir une quelconque signification, comme par exemple en français les mots « et », « est », « ou » ou encore « a ».

4.2 Motifs courts et motifs longs

Le théorème *Asymptotic Equipartition Property* de Shannon, McMillan et Breiman énonce que, pour une large classe de modèles probabilistes, la probabilité d'occurrence d'un motif typique de taille k est de l'ordre de 2^{-kh}. Le théorème assure de plus que l'ensemble des motifs du *mauvais ensemble* ou *bad set*, c'est-à-dire l'ensemble des motifs non-typiques, a une probabilité négligeable. Les motifs de taille plus petite que $c_1 \log n$ (qu'on a appelés courts dans le chapitre 3, c_1 est une constante inférieure à 1) et les motifs de taile supérieure à $c_2 \log n$ (appelés longs, c_2 est une constante supérieure à 1) sont non-typiques et ne comptent que très peu dans le comportement asymptotique. Cette section traite de la contribution asymptotique des motifs longs et courts à la taille moyenne de l'anti-dictionnaire. Les motifs qui ne sont ni longs, ni courts sont dits de taille intermédiaire (comme dans le chapitre 3) et l'ensemble des motifs de taille intermédiaire est noté $\mathfrak{I}_n$ (avec un indice n pour bien rappeler que la définition de ces motifs dépend de la taille du texte). La contribution des motifs intermédiaires est plus délicate à estimer. Les sections 4.3 à 4.5 sont dédiées à l'analyse de la contribution des motifs intermédiaires à la moyenne de la taille de l'anti-dictionnaire.

4.2.1 Motifs courts

Les motifs courts pour les textes de taille n sont définis (comme dans les sections 2.4.1 et 2.5.1) comme les motifs de taille inférieure à la borne

$$k_c(n) = (5/6)C_q \log n \text{ où } C_q = (-\log q)^{-1} \text{ et } p \geq q.$$

Nous montrons dans cette section que la contribution de ces motifs courts à la moyenne de la taille d'un anti-dictionnaire est très faible : en $o(1)$.

Pour que le motif w soit un MMI du texte T, il faut qu'il n'apparaisse pas dans le texte, c'est-à-dire que $\widehat{N}_w(T) = 0$. La probabilité que w soit un MMI est ainsi majorée par la probabilité que w n'apparaisse pas dans le texte. Les motifs de petites tailles ont une forte tendance à apparaître. Donc peu de textes vérifient $\widehat{N}_w(T) = 0$ pour w court. Notons qu'un MMI w doit avoir une taille k supérieure à 2 pour que ses préfixe et suffixe de taille $k-1$ ne soient pas réduits au mot vide ϵ. La contribution des motifs courts à la moyenne de la taille de l'anti-dictionnaire bâti sur un texte de taille n vérifie donc

$$\mathcal{F}(n) := \sum_{k=1}^{k_c(n)} \sum_{w \in \mathcal{A}^k} \mathbb{P}_n(w \in \text{ MMI}) \leq \sum_{k=2}^{k_c(n)} \sum_{w \in \mathcal{A}^k} \mathbb{P}_n(\widehat{N}_w = 0).$$

Un équivalent asymptotique de la probabilité que le motif w n'apparaisse pas dans un texte de taille n a été determiné par la proposition 5 de la section 2.3.3 page 47 :

$$\mathbb{P}_n(\widehat{N}_w = 0) \sim \exp\left(-\frac{np_w}{c(1)}\right).$$

Pour les motifs courts, np_w tend vers l'infini :

$$np_w \geq nq^k = n\exp(k \log q) \geq n \exp\left(\frac{5}{6} C_q \log n \log q\right) = n^{1-5/6}.$$

La valeur en 1 du polynôme d'auto-corrélation probabilisé est majorée par $(1-p)^{-1}$ où p est la plus grande des deux probabilités. La fonction $\exp(-x)$ est décroissante quand x tend vers

l'infini donc on majore chaque $\exp(-np_w/c(1))$ par sa plus grande valeur : $\exp((p-1)n^{1/6})$. La contribution des motifs courts est donc majorée

$$\mathcal{F}(n) \leq \sum_{k=2}^{k_c(n)} \sum_{w\in\mathcal{A}^k} \exp((p-1)n^{1/6}) = \sum_{k=2}^{k_c(n)} 2^k \exp((p-1)n^{1/6}) = Kn^{(5/6)C_q \log 2} \exp((p-1)n^{1/6}),$$

pour une constante positive K. La vitesse de décroissance vers zéro de la fonction exponentielle quand n tend vers l'infini permet d'énoncer le lemme :

Lemme 32 *La contribution des motifs courts à la moyenne sur les textes de taille n de la taille d'un anti-dictionnaire est en $o(1)$.*

Remarque : On a pris une définition des motifs courts avec un coefficient 5/6. Le résultat sur la contribution asymptotique des motifs courts à la moyenne de la taille d'un anti-dictionnaire est aussi valable pour toute borne $\alpha C_q \log n$ pourvu que α soit inférieur à 1.

4.2.2 Motifs longs

Les motifs longs pour les textes de taille n sont définis (cf. sections 2.4.2 et 2.5.2) comme les motifs de taille supérieure à

$$k_l(n) = 1.5C_p \log n \text{ où } C_p = (-\log p)^{-1}.$$

Nous établissons que la contribution $\mathcal{G}(n)$ des motifs longs à la taille moyenne, sur les textes de taille n, d'un anti-dictionnaire est en $O(n^{0.25})$.

Pour que le motif w soit un MMI du texte T, il faut nécessairement que $u = w_2 \cdots w_{k-1}$ apparaisse à au moins deux reprises dans le texte (une fois pour avoir l'occurrence de w_L et une autre fois pour l'occurrence de w_R). La probabilité qu'un motif w soit un MMI est majorée par la probabilité que $u(w)$ apparaisse à deux reprises, soit

$$\mathbb{P}_n(w \in \text{ MMI}) \leq \mathbb{P}_n(\widehat{N}_u \geq 2). \tag{4.4}$$

Or un motif long n'apparaît que très rarement à deux reprises dans un texte. Donc peu de textes vérifient $\widehat{N}_u(T) \geq 2$.

La contribution des motifs longs à la taille moyenne d'un anti-dictionnaire sur les textes de taille n se note

$$\mathcal{G}(n) := \sum_{k\geq k_l(n)} \sum_{w\in\mathcal{A}^k} \mathbb{P}_n(w \in \text{ MMI}). \tag{4.5}$$

Il existe quatre motifs w de taille k pour un motif u donné vérifiant $w_2 \cdots w_{k-1} = u$. L'équation (4.4) permet de majorer la contribution des motifs longs par

$$\mathcal{G}(n) \leq \sum_{k\geq k_l(n)} \sum_{w\in\mathcal{A}^k} \mathbb{P}_n(\widehat{N}_u \geq 2) = 4 \sum_{k\geq k_l(n)} \sum_{u\in\mathcal{A}^{k-2}} \mathbb{P}_n(\widehat{N}_u \geq 2).$$

Le terme dominant du comportement asymptotique de la probabilité d'avoir plus de deux occurrences d'un motif u a été obtenu dans les propositions 4 et 5 du chapitre 3 :

$$\begin{aligned} \mathbb{P}_n(\widehat{N}_u \geq 2) &= 1 - \mathbb{P}_n(\widehat{N}_u = 0) - \mathbb{P}_n(\widehat{N}_u = 1) \\ &\sim 1 - (1 + np_u) \exp\left(-\frac{np_u}{c_u(1)}\right). \end{aligned} \tag{4.6}$$

La taille de u est $k-2$ mais pour n grand la différence entre k et $k-2$ est négligeable. On utilisera donc abusivement l'entier k pour représenter la taille de u. Pour les motifs longs, np_u tend vers zéro quand n tend vers l'infini :

$$np_u \leq np^k \leq n\exp(1.5C_p \log n \log p) = \frac{1}{\sqrt{n}}. \tag{4.7}$$

Pour des x réels positifs, on a la majoration classique $\exp(-x) \geq 1-x$ et ainsi

$$1-(1+np_u)\exp\left(-\frac{np_u}{c_u(1)}\right) \leq np_u\left(\frac{1}{c(1)}-1+\frac{np_u}{c(1)}\right) = np_u f(u). \tag{4.8}$$

Nous sommes intéressés par la somme sur tous les motifs de taille k de $np_u f(u)$. Pour cela nous avons besoin de deux lemmes. Le lemme suivant permet de majorer ce type de somme :

Lemme 33 *Pour toute fonction f définie sur les motifs de $\mathcal{A}^\star$ et tout y,*

$$\sum_{w\in\mathcal{A}^k} p_w f(w) \leq \chi_k \mathbb{P}(\{w \in \mathcal{A}^k : f(w) > y\}) + y,$$

où χ_k est le maximum de la fonction f sur l'ensemble des motifs de taille k.

Preuve : La somme sur l'ensemble des motifs de taille k est divisée en deux sous-sommes : la somme sur les motifs pour lesquels $f(w) > y$ et la somme sur les autres motifs. La somme sur les motifs du premier ensemble est majorée par le produit de la plus grande valeur de f sur les motifs de taille k et la somme des probabilités des motifs de l'ensemble. Pour la somme sur les motifs du second ensemble, les valeurs de f sont, par définition de cet ensemble, majorées par y et la somme des probabilités de tous les motifs est majorée par 1. ◄

Le prochain lemme est adapté du lemme 7.6.4 de [JS05]. Il existe une version markovienne et étendue de ce lemme que l'on trouve dans le chapitre 5 (lemme 41 page 136).

Lemme 34 *Pour une source sans mémoire (p,q), tout entier k, $\theta = (1-p)^{-1}$ et $\delta = \sqrt{p}$,*

$$\sum_{w\in\mathcal{A}^k} [\![|c_w(1)-1| \leq \theta\delta^k]\!] p_w \geq 1-\theta\delta^k.$$

La preuve de ce lemme est donnée dans le chapitre 6. Pour un motif long u de taille k, on a

$$np_u \leq \left(p^{1/4}\right)^k.$$

Si $|c_u(1)-1| \leq \theta\delta^k$ alors $|c(1)-1|+np_u \leq \theta\delta^k + \left(p^{1/4}\right)^k$, ainsi en appliquant le lemme précédent

$$\sum_{u\in\mathcal{A}^k} [\![|f(u)| \leq \theta\delta^k + \left(p^{1/4}\right)^k]\!] p_u \geq \sum_{w\in\mathcal{A}^k} [\![|c_w(1)-1| \leq \theta\delta^k]\!] p_w \geq 1-\theta\delta^k.$$

Le y du lemme 33 est fixé à $\theta\delta^k + \left(p^{1/4}\right)^k$. La fonction $|f(u)|$ est majorée par $p + (\sqrt{n})^{-1}$ et il vient

$$\mathcal{G}(n) \leq 4n \sum_{k\geq k_l(n)} \sum_{u\in\mathcal{A}^k} p_u f(u) \leq 4n \sum_{k\geq k_l(n)} \left(p+\frac{1}{\sqrt{n}}+1\right)\theta\delta^k = O(n^{0.25}). \tag{4.9}$$

On peut ainsi énoncer le lemme :

Lemme 35 *La contribution $\mathcal{G}(n)$ des motifs longs (avec $k_l(n) = 1.5C_p \log n$) à l'espérance de la taille d'un anti-dictionnaire construit sur les textes de taille n se comporte asymptotiquement en $O(n^{0.25})$.*

Remarque : La contribution asymptotique des motifs longs dépend de la définition de la borne $k_l(n)$. Pour toute borne $k_l(n) = \beta C_p \log n$ avec $\beta > 1$, le comportement asymptotique de la contribution des motifs longs est sous-linéaire. Plus la constante β est proche de 1, plus la contribution asymptotique des motifs longs sera importante.

4.3 Présentation du modèle approché pour les motifs intermédiaires

Nous présentons le modèle de calcul approché sous lequel la contribution des motifs intermédiaires au comportement asymptotique de la taille moyenne d'un anti-dictionnaire sur les textes de taille n est quantifiée. Un motif intermédiaire est défini comme un motif w dont la taille k est comprise entre

$$(5/6)C_q \log n \quad \text{et} \quad 1.5C_p \log n \quad \text{où} \quad C_r = (-\log r)^{-1}.$$

L'ensemble des motifs intermédiaires relatifs aux textes de taille n est noté $\mathfrak{I}_n$. L'introduction de ce modèle vise à simplifier le calcul de la probabilité des textes de taille n pour lesquels w est un MMI. Le modèle approché est composé de deux hypothèses H_1 et H_2. Ces deux hypothèses sont ensuite utilisées pour approcher la contribution des motifs intermédiaires à la taille moyenne. Le modèle est validé dans les sections 4.6 et 4.7.

Pour que le motif w soit un MMI du texte T, il faut que les facteurs w_L et w_R apparaissent au moins une fois chacun dans le texte. Il faut donc au moins deux occurrences du motif $u := w_2 \cdots w_{k-1}$ (k est la taille de w) dans le texte. D'autre part, il est ardu de prendre en compte les possibilités de chevauchement entre les occurrences de w_L et w_R. De plus dans le chapitre 3, lors de l'étude de la taille et de la longueur de cheminement d'un arbre des suffixes, nous avons acquis l'intuition qu'un motif w va très rarement se chevaucher dans un texte. On introduit l'ensemble $\mathfrak{Y}_u$ des textes T avec au moins deux occurrences de u et pour lesquels toute occurrence de u est séparée d'une autre occurrence par au moins deux lettres. Dans un texte T appartenant à $\mathfrak{Y}_u$, le motif u apparaît à au moins deux reprises et les occurrences des *extensions* de u ne peuvent se chevaucher.

Hypothèse 1 : La probabilité des textes de taille n pour lesquels w est un MMI est approchée par la probabilité des textes de taille n avec au moins deux occurrences de u, dans lesquels deux occurrences quelconques de u sont séparées par au moins deux lettres (*i.e.,* des textes de $\mathfrak{Y}_u$) et pour lesquels w est un MMI. Autrement dit :

$$\mathbb{P}_n(w \in \text{ MMI}) \simeq \mathbb{P}(\{T : |T| = n,\ w \in \text{ MMI}(T) \text{ et } T \in \mathfrak{Y}_u\}) = \mathbb{P}_n(\{w \in \text{MMI}\} \cap \mathfrak{Y}_u) \quad (4.10)$$

Définition 13 *Le paramètre $\widetilde{N}_u$ est défini sur l'ensemble des textes. Si le texte T appartient à $\mathfrak{Y}_u$, le paramètre vaut le nombre d'occurrences du motif u dans le texte T. Si le texte T a moins de deux occurrences de u alors $\widetilde{N}_u(T) = N_u(T)$ et enfin si un texte T a au moins deux occurrences de u qui se chevauchent alors $\widetilde{N}_u(T) = 0$.*

Pour un texte T, un motif u et un entier $j \geq 2$, $\widetilde{N}_u(T) = j$ signifie que dans le texte T deux occurrences quelconques de u sont séparées par au moins 2 lettres et que le motif u apparaît à exactement j reprises dans le texte.

Hypothèse 2 : La loi du paramètre $\widetilde{N}_u$ sur les textes de taille n est approchée par une loi de Poisson de paramètre np_u. Cela s'écrit

$$\forall j \in \mathbb{N},\ \forall u \in \mathcal{A}^*, \quad \mathbb{P}_n(\widetilde{N}_u = j) \simeq \frac{(np_u)^j}{j!} \exp(-np_u). \tag{4.11}$$

4.3.1 Utilisation du modèle approché

Le modèle approché défini au-dessus est utilisé pour exprimer agréablement la probabilité qu'un motif de taille intermédiaire w soit un MMI. L'application de l'hypothèse H_1 à la probabilité que le motif de taille intermédiaire w soit un MMI entraîne

$$\mathbb{P}_n(w \in \text{ MMI}) \simeq \mathbb{P}_n(\{w \in \text{ MMI}\} \cap \mathfrak{Y}_u) = \sum_{j \geq 2} \mathbb{P}_n(\{w \in \text{ MMI}\} \cap \mathfrak{Y}_u \cap \{\widetilde{N}_u = j\}). \tag{4.12}$$

La somme porte sur les j supérieurs à 2 car le texte T appartient à $\mathfrak{Y}_u$. Puisque le motif w doit être de taille supérieure à 2, il se décompose sous la forme $\alpha u \beta$ où α et β sont deux lettres de l'alphabet. Pour chaque j, la probabilité $\mathbb{P}_n(\{w \in \text{ MMI}\} \cap \mathfrak{Y}_u \cap \{\widetilde{N}_u = j\})$ se récrit à l'aide de probabilités conditionnelles

$$\mathbb{P}_n(\{\alpha u \beta \in \text{ MMI}\} \cap \mathfrak{Y}_u \cap \{\widetilde{N}_u = j\}) = \mathbb{P}_n(\alpha u \beta \in \text{ MMI} \mid \mathfrak{Y}_u \cap \{\widetilde{N}_u = j\}) \mathbb{P}_n(\mathfrak{Y}_u \cap \{\widetilde{N}_u = j\}).$$

Si $\widetilde{N}_u(T)$ est supérieur à 2 alors le texte T appartient par définition à $\mathfrak{Y}_u$, ainsi pour $j \geq 2$

$$\mathfrak{Y}_u \cap \{\widetilde{N}_u = j\} = \{\widetilde{N}_u = j\}.$$

Sous le modèle approché (hypothèses H_1 et H_2), la contribution des motifs intermédiaires à la moyenne $\mathbb{E}_n(\mathcal{S})$ de la taille de l'anti-dictionnaire est approchée par

$$\mathcal{E}(n) := \sum_{k \in \mathfrak{I}_n} \sum_{\alpha,\beta \in \mathcal{A}^2} \sum_{u \in \mathcal{A}^{k-2}} \sum_{j \geq 2} \mathbb{P}_n(\alpha u \beta \in \text{ MMI } | \widetilde{N}_u = j) \frac{(np_u)^j}{j!} \exp(-np_u). \tag{4.13}$$

L'objectif des deux sections suivantes est de déterminer le comportement asymptotique de $\mathcal{E}(n)$ d'abord dans un modèle de source symétrique puis dans le cas d'une source biaisée. La proximité du modèle présenté dans cette section avec la réalité est validée dans les sections 4.6 et 4.7.

4.4 Source sans mémoire symétrique

Nous exprimons dans la sous-section 4.4.1, la probabilité des textes de taille n pour lesquels le mot w (de taille intermédiaire k, c'est-à-dire comprise entre $k_c(n)$ et $k_l(n)$) est un MMI sachant que les occurrences de u dans le texte sont séparées par au moins deux lettres. Le comportement asymptotique lorsque n tend vers l'infini de la contribution approchée $\mathcal{E}(n)$ des motifs intermédiaires à la moyenne $\mathbb{E}_n(\mathcal{S})$ sur les textes de taille n de la taille de l'anti-dictionnaire est obtenu en utilisant la théorie de Mellin. La somme $\mathcal{E}(n)$ se comporte asymptotiquement linéairement. Le coefficient de linéarité du comportement asymptotique de $\mathcal{E}(n)$ est déterminé explicitement.

Dans cette section, les textes T sont engendrés par une source sans mémoire symétrique (*i.e.* $p = q = 1/2$). C'est un cas plus facile à analyser que le cas biaisé qui sera traité dans la section suivante.

La transformée de Mellin ne peut s'appliquer directement sur la somme $\mathcal{E}(n)$ dont les indices de sommation dépendent du paramètre (la somme $\mathcal{E}(n)$ porte sur les tailles des motifs k entre $k_c(n)$ et $k_l(n)$). La stratégie adoptée ici consiste à étudier le comportement asymptotique de la somme $\Psi(n)$ du terme général de $\mathcal{E}(n)$ sur **tous** les indices k (section 4.4.2) et à ensuite montrer que les sommes des probabilités sur les motifs de taille inférieure à $k_c(n)$ ou de taille supérieure à $k_l(n)$ sont d'un ordre asymptotique sous-linéaire. Les sommes sur les motifs courts et longs sont majorées dans la sous-section 4.4.3. La somme sur les motifs courts se comporte asymptotiquement en $o(1)$ et la somme sur les motifs longs en $O(\sqrt{n})$. Ces deux sommes n'affectent donc pas le comportement asymptotique linéaire de la somme sur tous les indices et ce sont les motifs intermédiaires qui fournissent la contribution dominante à la somme.

4.4.1 Combinatoire des extrémités

Soit w un motif de taille k avec k compris entre $k_c(n)$ et $k_l(n)$, et $u = w_2 \dots w_{k-1}$. La lettre w_1 est notée α et la lettre w_k, β. Nous obtenons dans un premier temps la probabilité des textes de taille n pour lesquels w est un MMI et dans lesquels u apparaît à exactement deux reprises et où ces deux occurrences de u sont séparées par au moins 2 lettres. Ensuite nous traitons le cas général où le motif u apparaît un nombre j de fois, toutes les occurrences de u étant séparées par au moins deux lettres l'une de l'autre.

➜ La probabilité des textes de taille n pour lesquels w est un MMI, dans lesquels u apparaît à exactement deux reprises et ces deux occurrences soient séparées par au moins deux lettres se décompose en utilisant la probabilité conditionnelle :

$$\mathbb{P}_n(w \in \text{ MMI } \cap \widetilde{N}_u = 2) = \mathbb{P}_n(w \in \text{ MMI } | \widetilde{N}_u = 2)\mathbb{P}_n(\widetilde{N}_u = 2). \tag{4.14}$$

Sous l'hypothèse H_2 que nous avons introduite précédemment, la probabilité d'avoir exactement deux occurrences du motif u dans le texte et que ces deux occurrences soient séparées par au moins 2 lettres est approchée par

$$\mathbb{P}_n(\widetilde{N}_u = 2) \simeq \frac{(np_u)^2}{2!}\exp(-np_u) = \frac{1}{2!}\left(\frac{n}{2^{k-2}}\right)^2 \exp\left(-\frac{n}{2^{k-2}}\right).$$

Pour pouvoir garantir que w est un MMI, il faut s'assurer que chacune des deux occurrences de u ne doit créer avec ses deux lettres adjacentes (une à sa droite et une à sa gauche) aucune occurrence de w, tout en créant une occurrence de w_L et une de w_R. Les motifs w_L et w_R ne peuvent apparaître dans le texte qu'aux endroits où u apparaît (une occurrence de w_L ou w_R contient nécessairement une apparition de u). Nous avons précédemment introduit la notation $w = \alpha u \beta$. La lettre $\overline{\alpha}$ est la lettre de l'alphabet (binaire) qui n'est pas α. Nous avons besoin que l'une des occurrences de u soit encadrée par la paire de lettres adjacentes $(\alpha, \overline{\beta})$, ce qui crée une occurrence de w_L sans faire apparaître w. Nous avons aussi besoin que l'autre occurrence de u (le texte ne contient que deux occurrences de u) soit encadrée par la paire $(\overline{\alpha}, \beta)$ pour faire apparaître w_R sans avoir w.

Si nous comptons avec attention, il y a quatre lettres adjacentes à déterminer (les deux paires de lettres encadrant chacune des deux occurrences de u dans le texte). Puisque le modèle de source est symétrique, chacun des 16 choix possibles est équiprobable. D'autre part, nous avons deux choix quant à l'ordre d'apparition des paires de lettres adjacentes dans les textes pour

lesquelles w est un MMI (apparition de w_L avant ou après l'apparition de w_R). La probabilité d'avoir les lettres adjacentes qui permettent à w d'être un MMI lorsque u apparaît exactement à deux reprises dans le texte est donc 1/8. L'équation (4.19) étend ce résultat à un alphabet de taille quelconque et à un nombre quelconque d'apparitions du motif u (le nombre de lettres à choisir est toujours le double du nombre d'occurrences du motif u dans le texte).

Dans le cas particulier où le motif u apparaît à exactement deux reprises dans le texte de longueur n, la contribution approchée $F_2(n)$ des motifs intermédiaires à la taille moyenne s'exprime par la somme

$$\begin{aligned}\sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \mathbb{P}(w \in \text{ MMI } \cap \widetilde{N}_u = 2) &= \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \mathbb{P}(w \in \text{ MMI } |\widetilde{N}_u = 2)\mathbb{P}(\widetilde{N}_u = 2) \\ &\simeq \sum_{k=k_c(n)}^{k_l(n)} 2^k \frac{1}{2!8} \left(\frac{n}{2^{k-2}}\right)^2 \exp\left(-\frac{n}{2^{k-2}}\right) =: F_2(n).\end{aligned} \tag{4.15}$$

Remarque 1 : Nous avons besoin d'au moins deux occurrences de u pour que w puisse être un MMI : il faut une occurrence de w_L et une autre w_R, et que ces deux occurrences ne se partagent pas la même occurrence de u sans quoi w apparaît.
Remarque 2 : La taille du mot w doit être au moins 2 sinon w_L et w_R sont réduits au mot vide ϵ.

➜ Dorénavant, le motif u apparaît à exactement j reprises dans le texte avec $j \geq 2$ et chaque occurrence est séparée des autres par au moins deux lettres. Nous notons par $C_{j,u}$ la probabilité que les $2j$ lettres adjacentes aux j occurrences de u permettent à w d'être un MMI (on a toujours $2j$ lettres car le modèle exclut tout chevauchement). La probabilité $C_{j,u}$ est donc la probabilité que w soit un MMI sachant que u apparaît à exactement j reprises et que toute occurrence est séparée des autres par au moins deux lettres, soit

$$C_{j,u} := \mathbb{P}_n(w \in \text{ MMI } |\widetilde{N}_u = j) \tag{4.16}$$

Nous avons déjà vu en détail que $C_{2,u} = 1/8$ et ce pour tout motif u de taille supérieure à 2. Sous un modèle de source sans mémoire, $C_{j,u}$ ne dépend pas du motif u puisque les probabilités des lettres adjacentes à u valent nécessairement 0.5. La notation C_j est ainsi légitimée.

Les conditions pour que w soit un MMI sont, par définition, que w_L et w_R apparaissent à au moins une reprise chacun dans le texte et que w n'apparaisse jamais. Si l'on sait que le motif u apparaît à exactement j reprises dans le texte, cela revient à avoir au moins une paire de lettres adjacentes à une occurrence de u qui soit $(\alpha, \overline{\beta})$, au moins une autre paire de lettres adjacentes $(\overline{\alpha}, \beta)$ pour encadrer une occurrence de u et que jamais la paire (α, β) n'encadre une des j occurrences de u dans le texte. Pour ce qui est de la dernière des quatres paires de lettres adjacentes, $(\overline{\alpha}, \overline{\beta})$, elle peut apparaître un nombre quelconque de fois. L'hypothèse H_1 nous garantit que le choix des lettres adjacentes est libre puisque les occurrences de u sont séparées par au moins deux lettres. La série génératrice exponentielle (où z compte la longueur) des textes vérifiants cette propriété s'exprime

$$\sum_{j\geq 2} \frac{C_j}{j!} z^j = 1.(\exp(zp_\alpha p_{\overline{\beta}}) - 1)(\exp(zp_{\overline{\alpha}} p_\beta) - 1)\exp(zp_{\overline{\alpha}} p_{\overline{\beta}}) \tag{4.17}$$

Dans le cas d'une source symétrique sur un alphabet binaire, nous obtenons

$$C_j = \frac{1}{4^j}\left(3^j - 2^{j+1} + 1\right). \tag{4.18}$$

Le résultat s'étend à un alphabet de taille m

$$\begin{aligned} C_j &= j![z^j]\left[\exp\left(\frac{z(m-1)(m+1)}{4}\right) - 2\exp\left(\frac{zm(m-1)}{4}\right) + \exp\left(\frac{z(m-1)^2}{4}\right)\right] \\ &= \frac{1}{4^j}(m-1)^j\left[(m+1)^j - 2m^j + (m-1)^j\right]. \end{aligned} \tag{4.19}$$

En fait, il n'est pas nécessaire d'extraire le coefficient de la série génératrice exponentielle pour calculer la contribution approchée des motifs intermédiaires à la moyenne de la taille de l'anti-dictionnaire.

L'expression $F_j(n)$ de la contribution approchée (sous les hypothèses H_1 et H_2) des motifs intermédiaires à la taille moyenne de l'anti-dictionnaire quand le motif u apparaît à exactement j reprises et que ses occurrences sont séparées par au moins deux lettres entre elles s'écrit

$$\begin{aligned} \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \mathbb{P}_n(w \in \text{ MMI } \cap \widetilde{N}_u = j) &= \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \mathbb{P}_n(w \in \text{ MMI } | \widetilde{N}_u = j)\mathbb{P}_n(\widetilde{N}_u = j) \\ &\simeq \sum_{k=k_c(n)}^{k_l(n)} 2^k C_j \frac{1}{j!}\left(\frac{n}{2^{k-2}}\right)^j \exp\left(-\frac{n}{2^{k-2}}\right) =: F_j(n). \end{aligned}$$

Notons qu'il n'y a ni coefficient z^0, ni coefficient z dans cette série génératrice exponentielle. Ainsi, en sommant sur tout les nombres d'occurrences possibles de u (soit $j \geq 2$), nous aboutissons à la formule suivante pour $\mathcal{E}(n) = \sum_{j\geq 2} F_j(n)$

$$\begin{aligned} \mathcal{E}(n) &= \sum_{k=k_c(n)}^{k_l(n)} 2^k \left(\exp\left(\frac{n}{2^{k-2}}p_\alpha p_{\overline{\beta}}\right) - 1\right)\left(\exp\left(\frac{n}{2^{k-2}}p_{\overline{\alpha}}p_\beta\right) - 1\right)\exp\left(\frac{n}{2^{k-2}}p_{\overline{\alpha}}p_{\overline{\beta}}\right)\exp\left(-\frac{n}{2^{k-2}}\right) \\ &= \sum_{k=k_c(n)}^{k_l(n)} 2^k \left(\exp\left(\frac{n}{2^k}\right) - 1\right)^2 \exp\left(\frac{n}{2^k}\right)\exp\left(-\frac{n}{2^{k-2}}\right) \\ &= \sum_{k=k_c(n)}^{k_l(n)} 2^k \left(\exp\left(-\frac{n}{2^k}\right) - 2\exp\left(-\frac{2n}{2^k}\right) + \exp\left(-\frac{3n}{2^k}\right)\right). \end{aligned} \tag{4.20}$$

4.4.2 Analyse de Mellin

La transformée de Mellin ne permet pas d'étudier directement le comportement asymptotique de la somme $\mathcal{E}(n)$ quand n tend vers l'infini puisque la somme sur les indices k dépend du paramètre n. Pour palier cette difficulté, nous obtenons le comportement asymptotique de la somme $\Psi(n)$ sur les motifs w de toutes tailles de la probabilité sur les textes de taille n vérifiant l'hypothèse H_1 que w soit un MMI. La transformée de Mellin a déjà servi dans les chapitres précédents, les détails de l'analyse seront donc esquissés. Le comportement asymptotique de $\Psi(n)$ quand n tend vers l'infini est linéaire. Dans la section 4.4.3, on montre que les contributions des motifs courts et des motifs longs contribuent asymptotiquement sous-linéairement à la somme $\Psi(n)$ et qu'ainsi la contribution $\mathcal{E}(n)$ des motifs intermédiaires est linéaire.

La fonction

$$\Psi(z) := \sum_{k\geq 2} 2^k \left(\exp\left(-\frac{z}{2^k}\right) - 2\exp\left(-\frac{2z}{2^k}\right) + \exp\left(-\frac{3z}{2^k}\right)\right)$$

est à la base de notre étude puisque notre objectif est d'obtenir le comportement asymptotique de $\Psi(n)$. La fonction de base de la somme $\Psi(z)$ est

$$f(z) := \exp(-z) - 2\exp(-2z) + \exp(-3z).$$

La bande fondamentale de la fonction f est $\langle -2, \infty\rangle$. La bande de convergence de la série de Dirichlet $\sum_{k\geq 2} 2^{k(s+1)}$ de $\Psi^\star(s)$ (bande à l'intérieur de laquelle la somme converge) est $\langle -\infty, -1\rangle$. Ainsi la transformée de Mellin $\Psi^\star(s)$ est définie dans la bande $\langle -2, -1\rangle$ et y vaut

$$\Psi^\star(s) = \sum_{k\geq 2} 2^k \Gamma(s) 2^{ks}(1 - 2.2^{-s} + 3^{-s}) = \Gamma(s)\frac{2^{2(s+1)}}{1-2^{s+1}}(1 - 2.2^{-s} + 3^{-s}).$$

Seuls les pôles de $\Psi^\star(s)$ à droite de la bande $\langle -2, -1\rangle$ nous intéressent pour le comportement asymptotique de Ψ en l'infini. La transformée $\Psi^\star(s)$ a des pôles simples aux points $s = -1 + \frac{2im\pi}{\log 2}$ pour $m \in \mathbb{Z}^*$ puisqu'il y a un pôle simple pour le dénominateur en ces points. La quantité $1 - 2.2^{-s} + 3^{-s}$ ne peut s'annuler car le rapport $\log 3/\log 2$ est irrationnel. Les résidus en chacun de ces points valent

$$\Gamma\left(-1 + \frac{2im\pi}{\log 2}\right)\frac{1}{-\log 2}\left(1 - 2.2^{1-\frac{2im\pi}{\log 2}} + 3^{1-\frac{2im\pi}{\log 2}}\right) = -\frac{1}{\log 2}\Gamma\left(-1 + \frac{2im\pi}{\log 2}\right)\left(-3 + 3.3^{-\frac{2im\pi}{\log 2}}\right).$$

La transformée a aussi un pôle simple en $s = -1$ puisqu'elle a un pôle simple au dénominateur, un autre pôle simple pour la fonction Gamma et un zéro simple pour $1 - 2.2^{-s} + 3^{-s}$. Le résidu de la transformée de Mellin en -1 vaut

$$\frac{1}{\log 2}(4\log 2 - 3\log 3).$$

Nous sommons ensuite les résidus de chacun des pôles sis sur l'axe $\Re = -1$ et effectuons une transformation de Mellin « inverse » de manière à obtenir le comportement asymptotique de la taille de l'anti-dictionnaire quand les textes sont engendrés par une source sans mémoire symétrique. La fonction Gamma a un pôle en zéro mais celui-ci est effacé par le zéro de $1 - 2.2^{-s} + 3^{-s}$. Il n'y a donc pas d'autres pôles de la transformée de Mellin $\Psi^\star(s)$ à droite de $\Re = -1$. Le pôle en $s = -1$ apporte la plus large contribution au comportement asymptotique. La propriété de décroissance exponentielle de la fonction Gamma le long de l'axe imaginaire va rendre la contribution des autres pôles négligeable.

Lemme 36 *La somme $\Psi(n)$ a pour comportement asymptotique*

$$(3\log_2 3 - 4)n + \frac{n}{\log 2}\epsilon(n) + o(1), \tag{4.21}$$

où

$$\epsilon(n) := \sum_{m\in\mathbb{Z}^*} \Gamma\left(-1 + \frac{2im\pi}{\log 2}\right) 3\left(3^{-\frac{2im\pi}{\log 2}} - 1\right) n^{\frac{2im\pi}{\log 2}}$$

est une fonction oscillant autour de zéro de très faible amplitude (de l'ordre de 10^{-5}).

4.4.3 Majoration des sommes sur les motifs longs et courts

Dans cette section, nous montrons que les comportements asymptotiques de la somme sur les indices k inférieurs à $k_c(n)$ et de la somme sur les indices supérieurs à $k_l(n)$ de

$$2^k\left(\exp\left(-\frac{n}{2^k}\right)-2\exp\left(-\frac{2n}{2^k}\right)+\exp\left(-\frac{3n}{2^k}\right)\right)=2^k f\left(\frac{n}{2^k}\right)$$

sont sous-linéaires. Cela nous permet d'affirmer que le comportement asymptotique de la contribution approchée $\mathcal{E}(n)$ (sous les hypothèses H_1 et H_2) des motifs intermédiaires à la moyenne de la taille d'un anti-dictionnaire est linéaire.

➜ Motifs courts

Les arguments pour obtenir le comportement asymptotique de la somme des $2^k f(n/2^k)$ sur les k courts ont déjà été vus dans les sections 2.4.1 et 4.2.1. Puisque k est inférieur à $k_c(n)$, la quantité $n/2^k$ tend vers l'infini avec n. La quantité $n/2^k$ est minorée par $n^{1/6}$. La somme de $2^k f(n/2^k)$ sur les motifs courts est majorée par

$$\begin{aligned}\sum_{k=2}^{k_c(n)} 2^k f\left(\frac{n}{2^k}\right) &= \sum_{k=2}^{k_c(n)} 2^k \exp\left(-\frac{n}{2^k}\right)\left(1-2\exp\left(-\frac{n}{2^k}\right)+\exp\left(-\frac{2n}{2^k}\right)\right)\\ &\leq \exp(-n^{1/6})C\sum_{k=2}^{k_c(n)} 2^k = o(1),\end{aligned} \tag{4.22}$$

où C est une constante supérieure à 1 qui majore $1-2\exp\left(-\frac{n}{2^k}\right)+\exp\left(-\frac{2n}{2^k}\right)$ pour n suffisament grand (puisque cette quantité tend vers 1 quand n tend vers l'infini).

➜ Motifs longs

Les arguments pour obtenir le comportement asymptotique de la somme sur les longueurs supérieures à $k_l(n)$ ont déjà été rencontrés dans la section 2.4.2. La quantité $n/2^k$ tend vers zéro quand n tend vers l'infini. Le terme dominant du développement limité de $f(z)$ quand z tend vers zéro est z^2. Par conséquent, il vient l'approximation

$$\sum_{k\geq k_l(n)} 2^k f\left(\frac{n}{2^k}\right) \simeq \sum_{k\geq k_l(n)} 2^k\left(\frac{n}{2^k}\right)^2 = n^2\sum_{k\geq k_l(n)} 2^{-k} = O(\sqrt{n}). \tag{4.23}$$

La contribution asymptotique des k inférieurs à $k_c(n)$ à la somme $\Psi(n)$ est en $o(1)$ et la contribution asymptotique des k supérieurs à $k_l(n)$ est en $O(\sqrt{n})$. Donc le comportement asymptotique dominant de la somme de $f(n/2^k)$ sur les motifs intermédiaires est le même que le comportement asymptotique dominant de la somme $\Psi(n)$. Par conséquent

Proposition 6 *La contribution des motifs intermédiaires à la taille moyenne d'un anti-dictionnaire sur les textes de taille n engendrés par une source sans mémoire symétrique et sous les hypothèses H_1 et H_2 a pour comportement asymptotique*

$$\mathcal{E}(n) = (3\log_2 3-4)n + \frac{n}{\log 2}\epsilon(n) + O(\sqrt{n}), \tag{4.24}$$

où

$$\epsilon(n) := \sum_{m\in\mathbb{Z}^*} \Gamma\left(-1+\frac{2im\pi}{\log 2}\right) 3\left(3^{-\frac{2im\pi}{\log 2}}-1\right) n^{\frac{2im\pi}{\log 2}}$$

est une fonction oscillant autour de zéro de très faible amplitude (de l'ordre de 10^{-5}*).*

4.5 Source sans mémoire biaisée (p, q)

Dans cette section nous obtenons une expression de la contribution approchée (sous les hypothèses H_1 et H_2) des motifs intermédiaires à la moyenne sur les textes T de longueur n engendrés par une source biaisée (p, q) de la taille d'un anti-dictionnaire construit sur un texte T. Le comportement asymptotique quand n tend vers l'infini de cette contribution $\mathcal{E}(n)$ est ensuite déterminé : il est linéaire.

Comme dans le cas d'une source symétrique, plusieurs étapes sont nécessaires. D'abord la probabilité $P_{n,w}$ des textes de taille n pour lesquels w est un MMI sachant que les occurrences de u sont séparées par au moins deux lettres est déterminée. La théorie de Mellin permet d'obtenir le comportement asymptotique de la somme sur **tous** les motifs u des probabilités $P_{n,w}$. Ce comportement asymptotique est linéaire. Le comportement asymptotique de la somme des $P_{n,w}$ sur les motifs intermédiaires s'obtient en montrant que le comportement asymptotique des sommes sur les motifs courts et les motifs longs sont respectivement en $o(1)$ et en $O(\sqrt{n})$.

Nous avons vu dans l'équation (4.13) que la valeur approchée sous les hypothèses H_1 et H_2 de la contribution des motifs intermédaires à la moyenne de la taille de l'anti-dictionnaire sur les textes de taille n s'écrivait

$$\mathcal{E}(n) = \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^{k-2}} \sum_{\alpha,\beta\in\mathcal{A}^2} \sum_{j\geq 2} \mathbb{P}_n(\alpha u\beta \in \text{ MMI } |\widetilde{N}_u = j) \frac{(np_u)^j}{j!} \exp(-np_u).$$

Une remarque va nous simplifier les calculs : sous le modèle de génération des textes sans mémoire, la contribution du couple de lettres $\alpha = \mathtt{0}, \beta = \mathtt{1}$ est la même que la contribution du couple $\alpha = \mathtt{1}, \beta = \mathtt{0}$. En effet, la série génératrice exponentielle qui compte les probabilités des choix des extrémités et qui définit les coefficients

$$C_{j,u}^{\alpha\beta} := \mathbb{P}_n(\alpha u\beta \in \text{ MMI } |\widetilde{N}_u = j) \tag{4.25}$$

ne dépend ni de l'ordre des lettres, ni du motif u (c'est juste le choix des lettres extrêmes qui entre en compte). On peut donc omettre l'indice u du coefficient C_j. Le coefficient $C_j^{\mathtt{01}}$ est obtenu en extrayant le coefficient d'ordre j de la série génératrice exponentielle

$$\sum_{j\geq 0} \frac{C_j^{\mathtt{01}}}{j!} z^j = (\exp(p^2 z) - 1)(\exp(q^2 z) - 1)\exp(qpz). \tag{4.26}$$

Or on a la même série génératrice exponentielle pour le coefficient $C_j^{\mathtt{10}}$. Ainsi pour tout entier j, $C_j^{\mathtt{10}} = C_j^{\mathtt{01}}$.

Une autre difficulté s'ajoute à la multiplication du nombre de sommes (selon les valeurs de α et β) : la probabilité p_u d'occurrence du motif u ne dépend plus **que** de la taille du motif. Cette probabilité dépend aussi des lettres qui composent le motif (mais pas de l'ordre de ces lettres dans le motif). Puisque les sommes portent sur le motif u, il est plus pratique pour nos calculs de changer les indices de manière à ce que k soit la taille de u (au lieu de $k-2$). Il y a $\binom{k}{i}$ motifs

u de taille k avec i occurrences de la lettre 0 (et par conséquent $k-i$ occurrences de la lettre 1). Nous obtenons donc

$$\mathcal{E}(n) = \sum_{j\geq 2} \sum_{k=k_c(n)}^{k_l(n)} \sum_{\alpha,\beta\in\mathcal{A}^2} \sum_{i=0}^{k} \binom{k}{i} C_j^{\alpha\beta} \frac{(np^iq^{k-i})^j}{j!} \exp(-np^iq^{k-i}).$$

Le comportement asymptotique de cette somme est impossible à déterminer en se servant directement de la théorie de Mellin puisque l'indice k est sommé sur des valeurs qui dépendent du paramètre n. Comme dans le cas d'une source symétrique, nous regardons la somme sur toutes les longueurs k car la théorie de Mellin s'y applique. Nous définissons la somme

$$\begin{aligned}\Delta(n) &= \sum_{j\geq 2}\sum_{k\geq 0} \sum_{\alpha,\beta\in\mathcal{A}^2} \sum_{i=0}^{k} \binom{k}{i} C_j^{\alpha\beta} \frac{(np^iq^{k-i})^j}{j!} \exp(-np^iq^{k-i}) \\ &= \Delta^{00}(n) + \Delta^{11}(n) + \Delta^{10}(n) + \Delta^{01}(n),\end{aligned}$$

où les sommes $\Delta^{\alpha\beta}(n)$ sont indicées selon les lettres extrêmes de w.

4.5.1 $\Delta^{00}(z)$

La somme $\Delta^{00}(z)$ est la contribution des MMIs dont chacune des deux extrémités est la lettre 0 à la somme $\Delta(z)$. Soit

$$\begin{aligned}\Delta^{00}(z) :&= \sum_{k\geq 0}\sum_{i=0}^{k} \binom{k}{i} \Big[\exp(-zp^iq^{k-i})(\exp(zpqp^iq^{k-i}) - 1)^2 \exp(zq^2p^iq^{k-i})\Big] \\ &= \sum_{k\geq 0}\sum_{i=0}^{k} \binom{k}{i} \Big[\exp(zp^iq^{k-i}(-1+2pq+q^2)) - 2\exp(zp^iq^{k-i}(-1+pq+q^2)) \\ &\qquad + \exp(zp^iq^{k-i}(-1+q^2))\Big].\end{aligned}$$

La fonction de base associée à la fonction $\Delta^{00}(z)$ est

$$f(z) := \exp(z(-1+2pq+q^2)) - 2\exp(z(-1+pq+q^2)) + \exp(z(-1+q^2)).$$

La bande fondamentale de f est $\langle -2, -\infty\rangle$. La bande de convergence de la série de Dirichlet $\sum_{k\geq 0}(p^{-s}+q^{-s})^k$ est $s<-1$. Ainsi la transformée de Mellin de la fonction $\Delta^{00}(z)$ est définie à l'intérieur de la bande $\langle -2, -1\rangle$ et y vaut

$$\left(\Delta^{00}\right)^\star(s) = \Gamma(s)\frac{1}{1-(p^{-s}+q^{-s})}\Big[(1-2pq-q^2)^{-s} - 2(1-pq-q^2)^{-s} + (1-q^2)^{-s}\Big].$$

La transformée de Mellin a un pôle simple en -1 car $(1-2pq-q^2)^{-s}-2(1-pq-q^2)^{-s}+(1-q^2)^{-s}$ n'y a qu'un zéro simple. Le comportement asymptotique que ce pôle apporte est

$$\frac{1}{h}\big[(1-2pq-q^2)\log(1-2pq-q^2) - 2(1-pq-q^2)\log(1-pq-q^2) + (1-q^2)\log(1-q^2)\big]n.$$

Si la source est périodique (*i.e.* $\log p/\log q \in \mathbb{Q}$), il y a des pôles simples de la transformée de Mellin pour chaque $s_k := -1 + \frac{2ik\pi}{\log p - \log q}$ pour $k\in\mathbb{Z}^\star$. Les résidus en chacun des pôles simples s_k sont de la forme

$$\Gamma\left(-1+\frac{2ik\pi}{\log p-\log q}\right)\frac{1}{-p\log p - q\log q}\Big[p^{-2s_k} - 2p^{-s_k} + (1-q^2)^{-s_k}\Big],$$

et après utilisation du théorème de Mellin « inverse » leur contribution s'écrit

$$\frac{n}{h}\sum_{k\in\mathbb{Z}^*}\Gamma\left(-1+\frac{2ik\pi}{\log p-\log q}\right)\Big[p^{-2s_k}-2p^{-s_k}+(1-q^2)^{-s_k}\Big]n^{-s_k-1}.$$

La décroissance exponentielle de la fonction Gamma le long de l'axe imaginaire (ici $\Re(s)=-1$) fait que la contribution des pôles imaginaires au comportement asymptotique est négligeable et que seule la contribution du pôle en $s=-1$ va importer. Au final, le comportement asymptotique de la somme $\Delta^{00}(n)$ est

$$\begin{aligned}&\frac{n}{h}\big[p^2\log p^2-2p\log p+(1-q^2)\log(1-q^2)\big]\\&+\frac{n}{h}\sum_{k\in\mathbb{Z}^*}\Gamma(s_k)\Big[p^{-2s_k}-2p^{-s_k}+(1-q^2)^{-s_k}\Big]n^{-s_k-1}+o(n).\end{aligned}\tag{4.27}$$

Si la source est apériodique alors le comportement asymptotique de la somme est simplement

$$\frac{n}{h}\big[p^2\log p^2-2p\log p+(1-q^2)\log(1-q^2)\big]+o(n).\tag{4.28}$$

4.5.2 $\Delta^{11}(z)$

La fonction $\Delta^{11}(z)$ désigne la contribution des MMIs dont les extrémités sont, en première et en dernière position, 1. La méthode du paragraphe précédent permet de déterminer le comportement asymptotique de la fonction $\Delta^{11}(n)$.

$$\begin{aligned}\Delta^{11}(z)&:=\sum_{k\geq 0}\sum_{i=0}^{k}\binom{k}{i}\Big[\exp(-zp^iq^{k-i})(\exp(zpqp^iq^{k-i})-1)^2\exp(zp^2p^iq^{k-i})\Big]\\&=\sum_{k\geq 0}\sum_{i=0}^{k}\binom{k}{i}\Big[\exp(zp^iq^{k-i}(-1+2pq+p^2))-2\exp(zp^iq^{k-i}(-1+pq+p^2))\\&\quad+\exp(zp^iq^{k-i}(-1+p^2))\Big].\end{aligned}$$

La bande fondamentale de la transformée de Mellin est $\langle -2,\infty\rangle$ et la bande de convergence est $s<-1$. La transformée de Mellin est donc définie dans la bande $\langle -2,-1\rangle$ et y vaut

$$\left(\Delta^{11}\right)^\star(s):=\Gamma(s)\frac{1}{1-(p^{-s}+q^{-s})}\Big[(1-2pq-p^2)^{-s}-2(1-pq-p^2)^{-s}+(1-p^2)^{-s}\Big].$$

Cette transformée a un pôle simple en $s=-1$. Si la source est périodique, les $s_k:=-1+\frac{2ik\pi}{\log p-\log q}$ pour $k\in\mathbb{Z}^*$ sont des pôles simples. Le théorème de Mellin « inverse » donne le comportement asymptotique de la somme $\Delta^{11}(n)$ lorsque n tend vers l'infini

$$\begin{aligned}&\frac{n}{h}\big[q^2\log q^2-2q\log q+(1-p^2)\log(1-p^2)\big]\\&+\frac{n}{h}\sum_{k\in\mathbb{Z}^*}n^{-s_k-1}\Gamma(s_k)\Big[q^{-2s_k}-2q^{-s_k}+(1-p^2)^{-s_k}\Big]+o(n),\end{aligned}\tag{4.29}$$

si la source est périodique, et

$$\frac{n}{h}\big[q^2\log q^2-2q\log q+(1-p^2)\log(1-p^2)\big]+o(n)\tag{4.30}$$

sinon.

4.5.3 $\Delta^{\mathtt{10}}(z)$

La fonction $\Delta^{\mathtt{10}}(z)$ est la somme des contributions des MMIs qui commencent par la lettre 1 et finissent par la lettre 0. On utilise le même raisonnement que dans les paragraphes précédents.

$$\begin{aligned}\Delta^{\mathtt{10}}(z) &:= \sum_{k\geq 0}\sum_{i=0}^{k}\binom{k}{i}\Big[\exp(-zp^iq^{k-i})(\exp(zp^2p^iq^{k-i})-1)(\exp(zq^2p^iq^{k-i})-1)\exp(zpqp^iq^{k-i})\Big]\\ &= \sum_{k\geq 0}\sum_{i=0}^{k}\binom{k}{i}\Big[\exp(zp^iq^{k-i}(-1+p^2+q^2+pq)) - \exp(zp^{i-1}q^{k-i}(-1+pq+p^2))\\ &\quad - \exp(zp^iq^{k-i}(-1+pq+q^2)) + \exp(zp^iq^{k-i}(-1+pq))\Big].\end{aligned}$$

La bande fondamentale de la transformée de Mellin est $\langle -2, \infty\rangle$ et la bande de convergence $\langle -\infty, -1\rangle$. Ainsi la transformée de Mellin est définie dans la bande $\langle -2, -1\rangle$ et y vaut

$$\Gamma(s)\frac{1}{1-(p^{-s}+q^{-s})}\Big[(1-pq-p^2-q^2)^{-s} - (1-pq-p^2)^{-s} - (1-pq-q^2)^{-s} + (1-pq)^{-s}\Big].$$

Le comportement asymptotique de $\Delta^{\mathtt{10}}(n)$ quand n tend vers l'infini est

$$\begin{aligned}&\frac{n}{h}\big[pq\log pq - q\log q - p\log p + (1-pq)\log(1-pq)\big]\\ &+\frac{n}{h}\sum_{k\in\mathbb{Z}^*} n^{-s_k-1}\Gamma(s_k)\Big[(pq)^{-s_k} - p^{-s_k} - q^{-s_k} + (1-pq)^{-s_k}\Big] + o(n).\end{aligned} \tag{4.31}$$

et dans le cas d'une source apériodique

$$\frac{n}{h}\big[pq\log pq - q\log q - p\log p + (1-pq)\log(1-pq)\big] + o(n). \tag{4.32}$$

4.5.4 Majoration des sommes sur les motifs courts et longs

Nous montrons ici que pour chacune des quatre sommes $\Delta^{\alpha\beta}(n)$, les comportements asymptotiques de la somme sur les indices k inférieurs à $k_c(n)$ et de la somme sur les indices supérieurs à $k_l(n)$ sont sous-linéaires. Cela nous permet d'affirmer que le comportement asymptotique de la contribution approchée $\mathcal{E}(n)$ (sous les hypothèses H_1 et H_2) des motifs intermédiaires à la moyenne de la taille d'un anti-dictionnaire est linéaire. Les arguments pour justifier du comportement asymptotique des sommes sur les motifs courts et longs sont exposés uniquement pour la somme $\Delta^{\mathtt{00}}(n)$. Pour les trois autres sommes, la méthode et les résultats sont identiques.

➜ La contribution asymptotique des motifs courts à la somme $\Delta^{\mathtt{00}}(n)$ est en $o(1)$.

Nous rappelons d'abord la définition de la somme $\Delta^{\mathtt{00}}(n)$

$$\Delta^{\mathtt{00}}(n) = \sum_{k\geq 0}\sum_{i=0}^{k}\binom{k}{i}\Big[\exp(-np^iq^{k-i})(\exp(npqp^iq^{k-i})-1)^2\exp(nq^2p^iq^{k-i})\Big].$$

La quantité np^iq^{k-i} est supérieure à $n^{1/6}$ pour $|u| = k \leq k_c(n)$ et pour tout i entre 0 et k. Ainsi np^iq^{k-i} tend vers l'infini avec n pour tout motif court de taille k et avec i occurrences de la

lettre 0. Le terme général $\eta(n)$ de la somme se factorise

$$\begin{aligned}\eta(n) &:= \exp\left(-np^iq^{k-i}\right)\left(\exp\left(npqp^iq^{k-i}\right)-1\right)^2\exp\left(nq^2p^iq^{k-i}\right)\\ &= \exp\left(-np^iq^{k-i}(1-2pq-q^2)\right)\left(1-2\exp\left(-np^iq^{k-i}pq\right)+\exp\left(-np^iq^{k-i}2pq\right)\right).\end{aligned} \tag{4.33}$$

La quantité $1-2pq-q^2$ est positive donc la première exponentielle de la factorisation tend vers zéro quand n tend vers l'infini alors que le second facteur tend vers 1 quand n tend vers l'infini (donc ce terme est inférieur à une constante $C>1$ pour n assez grand). La somme sur les motifs de taille inférieure à $k_c(n)$ est donc majorée brutalement par

$$\sum_{k=0}^{k_c(n)} 2^kC\exp(-n^{1/6}(1-2pq-q^2)) = Kn^{-\frac{5}{6\log q}}\exp(-n^{1/6}(1-2pq-q^2)) = o(1),$$

pour une certaine constante K.

➜ **La contribution asymptotique des motifs longs à la somme $\Delta^{00}(n)$ est en $O(\sqrt{n})$.**

La quantité np^iq^{k-i} tend vers zéro quand n tend vers l'infini pour toute taille $k \geq k_l(n)$ et tout entier i entre 0 et k. Le terme dominant du développement limité du terme général $\eta(z)$ de la somme quand z est proche de zéro est $z^2(pqp_u)^2$. Cela permet d'obtenir l'approximation

$$\sum_{k\geq k_l(n)}\sum_{i=0}^{k}\binom{i}{n}\eta(n) \simeq \sum_{k\geq k_l(n)} n^2(pq)^2(p^2+q^2)^k = n^2(pq)^2(p^2+q^2)^{k_l(n)}.$$

La majoration de p^2+q^2 par p permet d'affirmer que le comportement asymptotique des motifs longs est en $O(\sqrt{n})$.

4.5.5 Addition des différentes contributions

Le comportement asymptotique de la somme $\Delta(n)$ est la somme sur tous les motifs est linéaire. La contribution asymptotique des motifs courts et des motifs longs à la somme $\Delta(n)$ est, au pire, en $O(\sqrt{n})$. Ainsi ce sont les motifs intermédiaires qui fournissent la contribution asymptotique linéaire de la somme $\Delta(n)$. Dans le cas d'une source périodique, nous sommons les équations (4.27), (4.31) et (4.29) et aboutissons à un comportement asymptotique pour la contribution des motifs intermédiaires à la valeur approchée de la taille moyenne d'un anti-dictionnaire sur les textes de taille n dans un modèle biaisé sans mémoire. Dans le cas d'une source apériodique, on somme les équations (4.28), (4.32) et (4.30) pour arriver au comportement asymptotique de la contribution des motifs intermédiaires à la taille moyenne de l'anti-dictionnaire.

Proposition 7 *Sous les hypothèses H_1 et H_2, la somme $\mathcal{E}(n)$ a pour comportement asymptotique*

$$\begin{aligned}&\frac{n}{h}[2h+(1-p^2)\log(1-p^2)+(1-q^2)\log(1-q^2)+2(1-pq)\log(1-pq)]\\ &+\frac{n}{h}\sum_{k\in\mathbb{Z}^*} n^{-s_k-1}\Gamma(s_k)\Big[\left[p^{-2s_k}-2p^{-s_k}+(1-q^2)^{-s_k}\right]+\left[q^{-2s_k}-2q^{-s_k}+(1-p^2)^{-s_k}\right]\\ &\quad+2\left[(pq)^{-s_k}-q^{-2s_k}-p^{-2s_k}+(1-pq)^{-s_k}\right]\Big]+o(n),\end{aligned}$$

dans le cas d'une source biaisée sans mémoire et périodique, et

$$\frac{n}{h}[2h + (1-p^2)\log(1-p^2) + (1-q^2)\log(1-q^2) + 2(1-pq)\log(1-pq)] + o(n),$$

pour une source biaisée sans mémoire et apériodique. La constante $h := -p\log p - q\log q$ *désigne l'entropie de la source.*

4.6 Validation de l'hypothèse H_1

La contribution des motifs intermédiaires à la taille moyenne d'un anti-dictionnaire sous le modèle approché défini à la section 4.3 a un comportement asymptotique linéaire. Nous validons dans cette section l'hypothèse H_1. Cela consiste à montrer que la différence entre les contributions des motifs intermédiaires à la moyenne de la taille de l'anti-dictionnaire sous le modèle approché et sous le modèle « véritable » se comporte asymptotiquement en $O(n^{1-c})$ pour un réel positif c.

La différence entre la contribution des motifs intermédiaires à la taille moyenne d'un anti-dictionnaire sur les textes de taille n et leur contribution sous le modèle approché est

$$\Xi(n) := \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} |\mathbb{P}_n(w \in \text{ MMI}) - \mathbb{P}_n(w \in \text{ MMI} \cap \mathfrak{Y}_u)|. \tag{4.34}$$

La différence entre les deux probabilités $\mathbb{P}_n(w \in \text{ MMI})$ et $\mathbb{P}_n(w \in \text{ MMI} \cap \mathfrak{Y}_u)$ est majorée par la probabilité de l'ensemble des textes de taille n de $\mathfrak{X}_u$. L'ensemble $\mathfrak{X}_u$ est l'ensemble des textes avec au moins deux occurrences de $u = w_2 \cdots w_{k-1}$ (puisqu'il faut au moins deux occurrences de u dans le texte T pour que w en soit un MMI) et tels que au moins deux de ces occurrences soient soit chevauchantes, soit *mitoyennes* (la dernière lettre de la première occurrence est suivie par la première lettre de la seconde occurrence de u), soit séparées par une seule lettre. Un texte T de $\mathfrak{X}_u$ est

- soit dans l'ensemble $\mathfrak{P}_u$ des textes pour lesquels deux occurrences de u sont mitoyennes ;
- soit dans l'ensemble $\mathfrak{Q}_u$ des textes pour lesquels deux occurrences de u sont séparées par une seule lettre ;
- soit dans l'ensemble $\mathfrak{R}_u$ des textes pour lesquels deux occurrences de u se chevauchent.

Cette décomposition n'est pas disjointe. Il est possible qu'un texte T de $\mathfrak{X}_u$ apparaisse dans plusieurs de ces ensembles. La probabilité des textes de taille n de $\mathfrak{X}_u$ est majorée par

$$\mathbb{P}_n(\mathfrak{X}_u) \leq \mathbb{P}_n(\mathfrak{P}_u) + \mathbb{P}_n(\mathfrak{Q}_u) + \mathbb{P}_n(\mathfrak{R}_u). \tag{4.35}$$

L'objectif est de trouver le comportement asymptotique de la somme $\Xi(n)$ donc la différence entre la taille de w et la taille de u est négligeable. On pose $|u| = k$. À chaque motif u de taille k il existe quatre motifs w de taille $k+2$ qui vérifient $w_2 \cdots w_{k+1} = u$. Ainsi

$$\begin{aligned}\Xi(n) &\leq 4 \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n(\mathfrak{X}_u) \\ &\leq 4 \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} (\mathbb{P}_n(\mathfrak{P}_u) + \mathbb{P}_n(\mathfrak{Q}_u) + \mathbb{P}_n(\mathfrak{R}_u)) = 4\left(\mathcal{P}(n) + \mathcal{Q}(n) + \mathcal{R}(n)\right),\end{aligned} \tag{4.36}$$

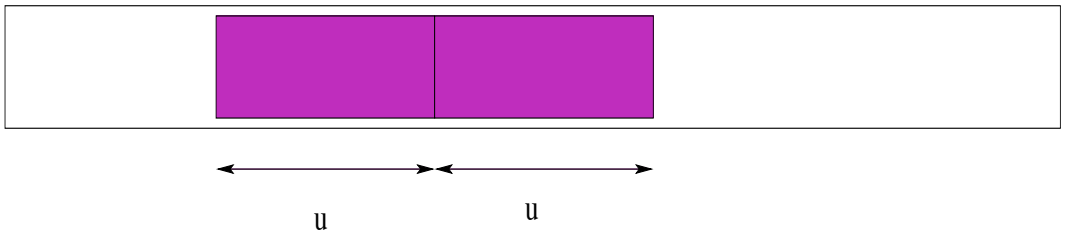


FIG. 4.2 – Deux occurrences de u mitoyennes dans un texte de $\mathfrak{P}_u$.

où

$$\mathcal{P}(n) := \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n(\mathfrak{P}_u),\ \mathcal{Q}(n) := \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n(\mathfrak{Q}_u) \text{ et } \mathcal{R}(n) := \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n(\mathfrak{R}_u).$$

Chacune de ces trois sommes est majorée séparément dans les sections suivantes selon une méthode adaptée.

4.6.1 Occurrences mitoyennes

La contribution asymptotique de la somme $\mathcal{P}(n)$ sur les motifs intermédiaires u de la probabilité des textes de $\mathfrak{P}_u$ de taille n est en $O(n^{1-\beta})$ où β est un réel positif.

Les textes de $\mathfrak{P}_u$ sont aussi les textes avec au moins une occurrence de $u.u$, soit les textes vérifiant $\widehat{N}_{uu} \geq 1$. L'application de la méthode du premier moment permet, à partir de la majoration

$$[\![\widehat{N}_{uu} \geq 1]\!] \leq \widehat{N}_{uu}[\![\widehat{N}_{uu} \geq 1]\!] \text{ d'obtenir } \mathbb{P}_n(\widehat{N}_{uu} \geq 1) \leq \mathbb{E}_n(\widehat{N}_{uu}).$$

La moyenne du nombre d'occurrences d'un motif v de taille l vaut $(n-l+1)p_v \leq np_v$. Nous obtenons ainsi la majoration

$$\mathbb{P}_n(\widehat{N}_{uu} \geq 1) \leq np_{uu} = np_u^2. \tag{4.37}$$

Et la somme $\mathcal{P}(n)$ est majorée par

$$\mathcal{P}(n) = \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n(\widehat{N}_{uu} \geq 1) \leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} np_u^2 \leq n \sum_{k=k_c(n)}^{k_l(n)} (p^2+q^2)^k = O\left(n\left(p^2+q^2\right)^{k_c(n)}\right).$$

La quantité p^2+q^2 est inférieure à 1 donc le rapport $\log(p^2+q^2)/\log q$ est positif. Ainsi l'exposant du comportement asymptotique est $1-\beta$ avec $\beta = (5/6)\log(p^2+q^2)/\log q$.

Lemme 37 *Le comportement asymptotique de la somme sur les motifs intermédiaires u de la probabilité des textes de taille n ayant deux occurrences mitoyennes de u est en $O(n^{1-\beta})$ où β est un réel positif.*

4.6.2 Occurrences séparées d'une lettre

La contribution asymptotique de la somme $\mathcal{Q}(n)$ sur les motifs intermédiaires de la probabilité des textes de taille n de $\mathfrak{Q}_u$ (textes avec au moins deux occurrences de u tels que deux de ces occurrences soient séparées par une seule lettre) est en $O(n^{1-\beta})$ où β est un réel positif.

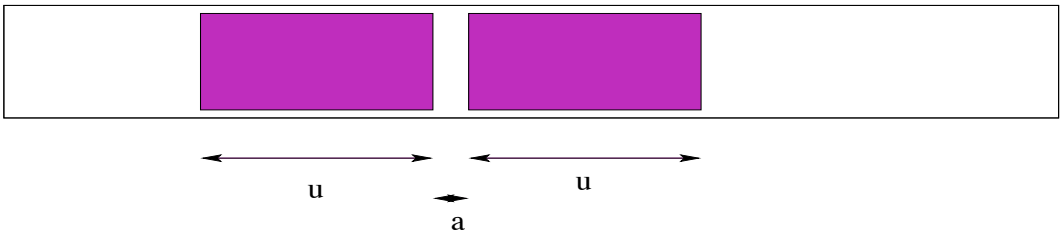


FIG. 4.3 – Deux occurrences de u séparées par une seule lettre ($a \in \mathcal{A}$) dans un texte de $\mathfrak{Q}_u$.

Les textes de $\mathfrak{Q}_u$ sont aussi décrits comme les textes avec au moins une occurrence de $u\alpha u$ pour $\alpha \in \mathcal{A}$, c'est-à-dire les textes vérifiant $\widehat{N}_{u0u} \geq 1$ ou $\widehat{N}_{u1u} \geq 1$. La majoration

$$[\![\widehat{N}_{u\alpha u} \geq 1]\!] \leq \widehat{N}_{u\alpha u}[\![\widehat{N}_{u\alpha u} \geq 1]\!]$$

est valable pour les deux lettres α de l'alphabet. Elle permet d'obtenir

$$\mathbb{P}_n(\widehat{N}_{u\alpha u} \geq 1) \leq \mathbb{E}_n(\widehat{N}_{u\alpha u}) \leq np_{u\alpha u} = np_u^2 p_\alpha.$$

La contribution de la somme sur les motifs intermédiaires des probabilités des textes de $\mathfrak{Q}_u$ est majorée par

$$\begin{aligned}\mathcal{Q}(n) &\leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \sum_{\alpha\in\mathcal{A}} \mathbb{P}_n(\widehat{N}_{u\alpha u} \geq 1) \leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \sum_{\alpha\in\mathcal{A}} np_u^2 p_\alpha \\ &\leq n \sum_{k=k_c(n)}^{k_l(n)} (p^2+q^2)^k = O\left(n\left(p^2+q^2\right)^{k_c(n)}\right).\end{aligned}$$

La quantité p^2+q^2 est inférieure à 1 donc le rapport $\log(p^2+q^2)/\log q$ est positif. Ainsi l'exposant du comportement asymptotique est $1-\beta$ avec $\beta = (5/6)\log(p^2+q^2)/\log q$.

Lemme 38 *Le comportement asymptotique de la somme sur les motifs u de taille intermédiaire apparaissant dans les textes de $\mathfrak{Q}_u$ de taille n est en $O(n^{1-\beta})$ où β est un réel positif.*

4.6.3 Occurrences chevauchantes

La contribution asymptotique de la somme $\mathcal{R}(n)$ sur les motifs intermédiaires u de la probabilité des textes de taille n de $\mathfrak{R}_u$ est en $O(n^{1-\delta})$ où δ est un réel positif.

L'ensemble des textes $\mathfrak{R}_u$ avec deux occurrences chevauchantes de u est subdivisé en $\mathfrak{R}_u^{(1)}$, l'ensemble des textes dans lesquels deux occurrences de u se chevauchent sur moins de la moitié de u (chevauchement « faible ») et l'ensemble $\mathfrak{R}_u^{(2)}$ des textes dans lesquels deux occurrences de u se chevauchent sur plus de la moitié du motif u (chevauchement « fort »). Les deux ensembles $\mathfrak{R}_u^{(1)}$ et $\mathfrak{R}_u^{(2)}$ ne sont pas disjoints puisqu'on peut avoir plusieurs couples de deux occurrences dans un texte. Donc

$$\mathbb{P}_n(\mathfrak{R}_u) \leq \mathbb{P}_n\left(\mathfrak{R}_u^{(1)}\right) + \mathbb{P}_n\left(\mathfrak{R}_u^{(2)}\right).$$

On introduit les sommes relatives à ces deux nouveaux ensembles :

$$\mathcal{R}^{(1)}(n) := \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n\left(\mathfrak{R}_u^{(1)}\right) \text{ et } \mathcal{R}^{(2)}(n) := \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n\left(\mathfrak{R}_u^{(2)}\right). \tag{4.38}$$

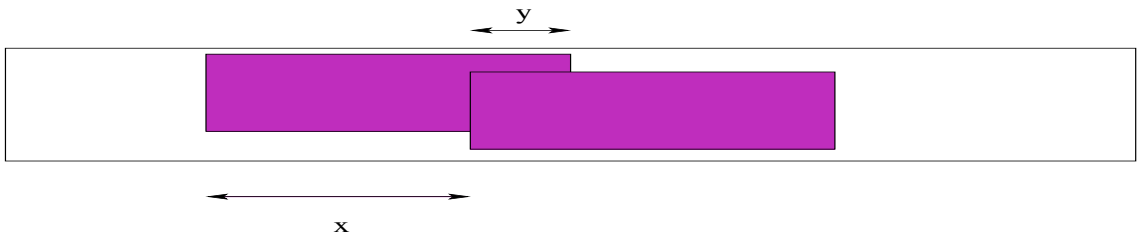


FIG. 4.4 – Deux motifs u se chevauchant dans le texte (un texte de $\mathfrak{R}_u$), x le préfixe de taille i de u où i est la distance entre les premières lettres de chacune des deux occurrences.

➜ Occurrences se chevauchant de manière « faible »

Les deux motifs u apparaissent dans le texte en se chevauchant faiblement. La distance entre les positions de la première lettre de chacune des deux occurrences de ce mot est notée i. Par définition des textes de $\mathfrak{R}_u^{(1)}$, l'entier i est supérieur à $k/2$ où k est la taille du motif u. Le préfixe de taille i de u est noté x. Les deux motifs u se chevauchant peuvent se récrire sous la forme xxy où $|y| < |x|$ et y est le préfixe de taille $k-i$ de x (c'est-à-dire que y est uniquement déterminé par x). La majoration utilisée dans l'équation (4.37) permet d'écrire

$$\begin{aligned}\mathcal{R}^{(1)}(n) &= \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n\left(\mathfrak{R}_u^{(1)}\right) \leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{i=\frac{k}{2}}^{k-1} \sum_{x\in\mathcal{A}^i} \mathbb{P}_n(\widehat{N}_{xxy(x)} \geq 1) \\ &\leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{i=\frac{k}{2}}^{k-1} \sum_{x\in\mathcal{A}^i} np_{y(x)}p_x^2 \leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{i=\frac{k}{2}}^{k-1} \sum_{x\in\mathcal{A}^i} np_x^2 = \sum_{k=k_c(n)}^{k_l(n)} \sum_{i=\frac{k}{2}}^{k-1} n(p^2+q^2)^i \\ &= O\left(n(p^2+q^2)^{k_c(n)/2}\right).\end{aligned} \tag{4.39}$$

Le comportement asymptotique de ce type de somme a déjà été traîté dans les deux sous-sections antérieures.

Lemme 39 *La contribution asymptotique de la somme sur les motifs u de taille intermédiaire des probabilités des textes de taille n dans lesquels deux occurrences de u apparaissent en se chevauchant sur moins de la moitié du motif est en $O(n^{1-\beta/2})$ pour un réel β positif.*

Remarque : Le fait que deux occurrences de u se chevauchent faiblement dans un texte (un texte de $\mathfrak{R}_u^{(1)}$ donc) ne contredit pas le fait que le motif u puisse être fortement corrélé (c'est-à-dire que l'indice positif du plus petit coefficient non nul du polynôme d'auto-corrélation du motif u peut être très faible).

➜ Occurrences se chevauchant de manière « forte »

Les occurrences de u se chevauchant dans le texte de manière forte sont des motifs avec une « forte » corrélation, c'est-à-dire qu'ils peuvent se chevaucher sur un large suffixe du motif. Ce type de motif est rare et la somme sur ces motifs de la probabilité des textes de taille n « fortement » chevauchants est en $O(n^{1-\delta})$ pour un réel $\delta > 0$.

Pour des motifs fortement corrélés, le plus petit degré positif des termes non nuls du polynôme d'auto-corrélation est inférieur à $k/2$. L'équation (5.7) (avec $z = 1$ et $j = k/2$) dans la preuve

du lemme 41 du chapitre 6 permet d'écrire que la somme des probabilités des motifs fortement corrélés décroit exponentiellement en fonction de la taille

$$\sum_{w\in\mathcal{A}^k} [\![c_w(1)-1 \text{ a degré minimal } \leq k/2]\!] p_w \leq \frac{p^{k/2}}{1-p}.$$

Un texte appartient à l'ensemble $\mathfrak{R}_u^{(2)}$ si deux occurrences de u se chevauchant fortement y apparaissent. Or pour que deux occurrences de u se chevauchent fortement dans un texte, il faut que le motif u soit fortement corrélé.

La probabilité que la configuration formée de deux occurrences de u se chevauchant fortement apparaisse dans le texte est majorée par la probabilité que le mot u apparaisse. Cette majoration est fine car elle revient à majorer la probabilité d'un suffixe de u de petite taille (donc de probabilité proche de 1) par 1. La somme sur les motifs de taille k est réduite à une somme sur les motifs de taille k qui sont fortement corrélés :

$$\begin{aligned}\mathcal{R}^{(2)}(n) &= \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n\left(\mathfrak{R}_u^{(2)}\right) \leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} [\![u \text{ fortement corrélé }]\!] \mathbb{P}_n(\widehat{N}_u \geq 1) \\ &\leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} [\![u \text{ fortement corrélé }]\!] np_u \leq \sum_{k=k_c(n)}^{k_l(n)} n\frac{p^{k/2}}{1-p} = O\left(np^{k_c(n)/2}\right).\end{aligned} \tag{4.40}$$

La puissance $p^{k_c(n)/2}$ se récrit sous la forme $n^{-\delta}$ avec $\delta = (5/12)(\log p/\log q)$. Or les probabilités p et q sont toutes deux plus petites que 1, donc $\delta > 0$.

Lemme 40 *Le comportement asymptotique de la somme sur les motifs de taille intermédiaire u de la probabilité des textes de taille n dans lesquels deux occurrences de u apparaissent en se chevauchant fortement est en $O(n^{1-\delta})$ avec $\delta > 0$.*

4.6.4 Conclusion

Sur les quatre types de textes étudiés dans les sections précédentes (textes de $\mathfrak{P}_u$, $\mathfrak{Q}_u$, $\mathfrak{R}_u^{(1)}$ et $\mathfrak{R}_u^{(2)}$), la plus grande contribution asymptotique vient de la somme sur les motifs intermédiaires de la probabilité des textes avec un chevauchement fort (*i.e.* des textes de $\mathfrak{R}_u^{(2)}$). Cela est du à l'inégalité $p^2+q^2 < p$ pour toute probabilité p entre 0.5 et 1.

Proposition 8 *La différence $\Xi(n)$ entre la contribution des motifs intermédiaires à l'espérance de la taille de l'anti-dictionnaire sur les textes de taille n et leur contribution à l'espérance de la taille d'un anti-dictionnaire sous le modèle approché vérifie asymtotiquement*

$$\Xi(n) = O\left(n^{1-\frac{5}{12}\frac{\log p}{\log q}}\right).$$

Le tableau suivant montre le comportement de l'exposant $1-\frac{5}{12}\frac{\log p}{\log q}$ selon les valeurs de la probabilité p. Plus la probabilité p est proche de 1, plus l'exposant va s'approcher de 1.

p	exposant (valeur approchée)
0.5	0.5833
0.6	0.7677
0.7	0.8765
0.8	0.9422
0.9	0.9809

4.7 Validation de l'hypothèse H_2

Dans cette section nous validons l'hypothèse H_2. L'hypothèse H_2 approche les valeurs de $\mathbb{P}_n(\widetilde{N}_u = j)$ pour $j \geq 2$ par les valeurs d'une loi de Poisson de paramètre np_u. Pour valider cette hypothèse, nous montrons que la différence entre la contribution « réelle » des motifs intermédiaires à la taille moyenne de l'anti-dictionnaire sur les textes de taille n et leur contribution sous le modèle approché se comporte asymptotiquement en $O(n^{1-\beta})$ avec un réel β positif.

Montrer que l'hypothèse H_2 est valide, revient à montrer que pour une variable aléatoire X_u suivant une loi de Poisson de paramètre np_u, la somme

$$\mathfrak{S}(n) := \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \sum_{j\geq 2} \mathbb{P}_n(w \in \text{ MMI } |\widetilde{N}_u = j) \left[\mathbb{P}_n(\widetilde{N}_u = j) - \mathbb{P}_n(X_u = j)\right] = o(n).$$

Notre résultat est plus fort puisque l'on montre que la somme $\mathfrak{S}(n)$ se comporte asymptotiquement en $O(n^{1-\beta})$ pour un $\beta > 0$.

La somme $\mathfrak{S}(n)$ se décompose en deux sous-sommes

- une somme $\mathfrak{S}_1(n)$ qui estime une « distance » entre la loi de Poisson et le paramètre N_u (qui compte le nombre d'occurrences du motif u dans le texte) ;
- une somme $\mathfrak{S}_2(n)$ qui estime une « distance » entre les paramètres N_u et $\widetilde{N}_u$.

Rappelons la notation $C_{j,u}^{\alpha,\beta} = \mathbb{P}_n(\alpha u\beta \in \text{ MMI } |\widetilde{N}_u = j)$. Les sommes $\mathfrak{S}_1(n)$ et $\mathfrak{S}_2(n)$ sont définies par

$$\begin{aligned}
\mathfrak{S}_1(n) &:= \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \sum_{j\geq 2} C_{j,u}^{\alpha,\beta} \left[\mathbb{P}_n(N_u = j) - \frac{(np_u)^j}{j!}\exp(-np_u)\right] \text{ et} \\
\mathfrak{S}_2(n) &:= \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \sum_{j\geq 2} C_{j,u}^{\alpha,\beta} \left[\mathbb{P}_n(N_u = j) - \mathbb{P}(\widetilde{N}_u = j)\right].
\end{aligned} \tag{4.41}$$

La première somme va être majorée dans la première sous-section en utilisant un théorème de Barbour, Holst et Janson [BHJ92] découlant de l'application de la méthode de Stein-Chen. Dans la seconde sous-section, la différence entre les probabilités $\mathbb{P}_n(\widetilde{N}_u = j)$ et $\mathbb{P}_n(N_u = j)$ sera réinterprétée pour majorer la somme $\mathfrak{S}_2(n)$. Les résultat de la section 4.6 permettent de conclure.

4.7.1 Majoration de la somme $\mathfrak{S}_1(n)$

Dans cette section, nous montrons que la somme $\mathfrak{S}_1(n)$ se comporte asymptotiquement en $O(\log^2(n))$.

Le théorème 8.F (p. 163) de [BHJ92] apporte la conclusion

$$\sum_{j\geq 0} \left|\mathbb{P}_n(N_u = j) - \frac{(np_u)^j}{j!}\exp(-np_u)\right| \leq 4(c_u(1) - 1) + 2(2|u| - 1)p_u. \tag{4.42}$$

Cette somme représente une distance entre la loi de Poisson de paramètre np_u et la loi de N_u.

Les coefficients $C_{j,u}^{\alpha,\beta}$ sont tous trivialement et grossièrement majorés par 2. Avec le modèle de source sans mémoire, le coefficient ne dépend pas du motif u mais uniquement des lettres extrêmes. Pour les extrémités `0` et `1`, le coefficient C_j s'obtient en extrayant le coefficient d'ordre j de la série génératrice exponentielle exprimée dans l'équation (4.26). La probabilité C_j vaut

alors $(1-pq)^j - p^j - q^j + (pq)^j \leq 2$. Un motif w ayant les extrémités `1` et `0` donne la même expression pour la probabilité C_j. Pour les motifs ayant deux extrémités `1`, le coefficient C_j s'écrit $(1-q^2)^j - 2p^j + p^{2j} \leq 2$. Quand les extremités de w sont toutes deux `0`, $C_j = (1-p^2)^j - 2q^j + q^{2j} \leq 2$.

L'étape suivante consiste à sommer sur tous les motifs intermédiaires

$$\begin{aligned}
|\mathfrak{S}_1(n)| &\leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \sum_{j\geq 2} C_{j,u} \left| \mathbb{P}_n(N_u = j) - \frac{(np_u)^j}{j!} \exp(-np_u) \right| \\
&\leq 2 \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} \sum_{j\geq 0} \left| \mathbb{P}_n(N_u = j) - \frac{(np_u)^j}{j!} \exp(-np_u) \right| \qquad (4.43) \\
&\leq 2 \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} (4(c_u(1)-1) + 2(2|u|-1)p_u).
\end{aligned}$$

Pour chaque motif u, il existe 4 motifs w possibles, selon les 4 choix possibles d'extrémités de w.

Le lemme 2.10 du chapitre 3 affirme que pour un modèle de source sans mémoire et pour toute longueur k,

$$\sum_{u\in\mathcal{A}^k} c_u(1) = 2^k + k - 1.$$

Il en résulte la majoration

$$|\mathfrak{S}_1(n)| \leq 16 \sum_{k=k_c(n)}^{k_l(n)} 2(k-1) + (2k-1) \leq 32 k_l^2(n) = O(\log^2(n)).$$

4.7.2 Majoration de la somme $\mathfrak{S}_2(n)$

Dans cette section, nous montrons que la somme $\mathfrak{S}_2(n)$ se comporte asymptotiquement en $O(n^{1-\beta})$ avec une constante positive β.

Soit un texte T ayant exactement $j \geq 2$ occurrences de u et que chacune est séparée des autres par au moins deux lettres. Ce texte vérifie nécessairement $N_u(T) = j$. Ainsi pour $j \geq 2$, la probabilité des textes vérifiant $N_u = j$ est toujours plus grande que celle des textes vérifiant $\widetilde{N}_u = j$. Cette remarque cesse d'être vraie pour $j = 1$ et 0. On écrit

$$\sum_{j\geq 2} \left| \mathbb{P}_n(N_u = j) - \mathbb{P}_n(\widetilde{N}_u = j) \right| = \mathbb{P}_n(\mathfrak{X}_u), \qquad (4.44)$$

où, comme dans la section 4.6, $\mathfrak{X}_u$ est l'ensemble des textes avec au moins deux occurrences de u et pour lesquels au moins deux de ces occurrences de u sont séparées par au plus une lettre (c'est-à-dire soit des occurrences chevauchantes, soit des occurrences mitoyennes, soit des occurrences séparées par une seule lettre).

Les coefficients $C_{j,u}^{\alpha,\beta}$ dans la somme $\mathfrak{S}_2(n)$ sont tous majorés par 2 comme au paragraphe précédent. Ainsi

$$|\mathfrak{S}_2(n)| \leq \sum_{k=k_c(n)}^{k_l(n)} \sum_{w\in\mathcal{A}^k} 2 \sum_{j\geq 2} \left| \mathbb{P}_n(N_u = j) - \mathbb{P}_n(\widetilde{N}_u = j) \right| \leq 8 \sum_{k=k_c(n)}^{k_l(n)} \sum_{u\in\mathcal{A}^k} \mathbb{P}_n(\mathfrak{X}_u). \qquad (4.45)$$

Le théorème 8 permet d'obtenir que la somme $\mathfrak{S}_2(n)$ se comporte asymptotiquement en $O(n^{1-\beta})$ avec β réel positif dépendant du modèle probabiliste sur la génération des textes. Avec les bornes sur les motifs intermédiaires $k_c(n) = (5/6)C_q \log n$ et $k_l(n) = 1.5C_p \log n$, la constante β vaut $(5/12) \log p / \log q$.

4.7.3 Conclusion

L'approximation effectuée par l'hypothèse H_2 consiste à dire que le nombre d'occurrences du motif u dans les textes de $\mathfrak{Y}_u$ de taille n suit une loi de Poisson de paramètre np_u.

Proposition 9 *Pour un modèle probabiliste sans mémoire biaisé de génération des textes, la différence $\mathfrak{S}(n)$ entre la contribution approchée des motifs intermédiaires à la taille moyenne de l'anti-dictionnaire sous l'hypothèse H_1 et la contribution approchée en utilisant les hypothèses H_1 et H_2 est*

$$\mathfrak{S}(n) = O\left(n^{1-\frac{5\log p}{12\log q}}\right).$$

Puisque la somme est sous-linéaire, l'hypothèse H_2 et validée.

4.8 Conclusion

La section 4.2 a permis de montrer que la contribution des motifs courts à la moyenne de la taille d'un anti-dictionnaire est asymptotiquement en $o(1)$ et celle des motifs longs en $O(n^\alpha)$ (où l'exposant $\alpha < 1$ du comportement asymptotique des motifs longs dépend du choix de $k_l(n)$). Dans la section 4.5, nous avons montré que le comportement asymptotique de la contribution des motifs intermédiaires sous le modèle approché de la section 4.3 est linéaire (et le coefficient de linéarité est connu). La différence entre la contribution des motifs intermédiaires à la moyenne de l'anti-dictionnaire et leur contribution sous le modèle approché est en $O(n^{1-\beta})$ (où $\beta > 0$ dépend des bornes $k_c(n)$ et $k_l(n)$). Nous avons obtenu le théorème :

Théorème *Soit un modèle probabiliste sans mémoire biaisé (p, q) de génération des textes. Le comportement asymptotique de la moyenne $\mathbb{E}_n(\mathcal{S})$ sur les textes de taille n de la taille de l'anti-dictionnaire vaut dans le cas d'une source périodique*

$$\mathbb{E}_n(\mathcal{S}) = K\frac{n}{h} + \frac{n}{h}\epsilon(n) + o(n) \tag{4.46}$$

et dans le cas d'une source apériodique,

$$\mathbb{E}_n(\mathcal{S}) = K\frac{n}{h} + o(n), \tag{4.47}$$

$$\text{où} \quad K := 2h + (1-p^2)\log(1-p^2) + (1-q^2)\log(1-q^2) + 2(1-pq)\log(1-pq)$$

et $\epsilon(n)$ est une fonction oscillant autour de zéro de très faible amplitude.

Ota Takahiro de l'*Institut préfectoral de technologie* de Nagano a effectué des simulations sur le nombre moyen de mots dans un anti-dictionnaire quand la taille du texte est de 3000 caractères. Nous comparons ses simulations au résultat du théorème précédent en considérant qu'un texte de taille 3000 a une longueur asymptotique.

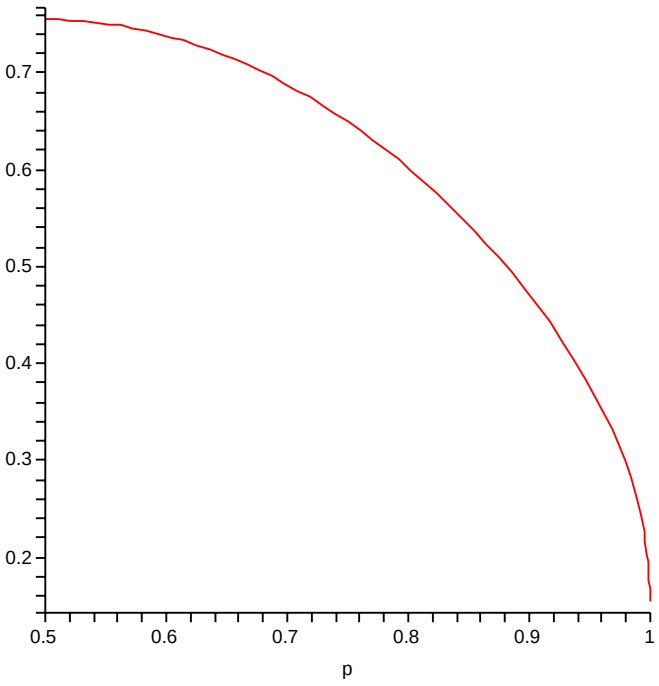


FIG. 4.5 – Comportement de la constante de linéarité en fonction de la probabilité p.

p	Simulations	Modèle approché
0.5	2273	2264
0.6	2213	2214
0.7	2076	2063
0.8	1811	1804
0.9	1346	1412

Quand la probabilité p est proche de 1, les simulations perdent un peu de leur précision. L'erreur entre les simulations et le résultat sous modèle approché sont au plus de 0.6 % pour les probabilités inférieures à 0.9 et de 5 % sinon. Ces pourcentages d'erreur sont très satisfaisants vis-à-vis de la validité de notre modèle approché.

Plusieurs axes de recherche s'ouvrent pour améliorer ce travail. D'abord approfondir nos connaissance sur le nombre de mots dans l'anti-dictionnaire d'un texte en regardant leur variance et leur distribution. L'étude (moyenne, variance, distribution) de la **longueur** des mots dans l'anti-dictionnaire permet d'affiner la connaissance du comportement asymptotique de l'algorithme `DCA` et en particulier d'obtenir le taux de compression de l'algorithme. Dans [CMRS99], Crochemore *et alii* proposent un anti-dictionnaire modifié qui ne contient que les mots de taille plus petite qu'une borne donnée. Le comportement de la longueur des MMIs permet alors de mieux connaître les paramètres de ce type d'anti-dictionnaire et aussi de bien choisir la borne sur les longueurs.

Les résultats expérimentaux de [CMRS99] mettent en évidence un taux de compression pour une version particulière de l'algorithme `DCA` qui s'approche de l'entropie. Leurs simulations sont validées par un résultat théorique (qui utilise la théorie des automates) dans le cas d'une

source symétrique. Il nous faudra expliquer, à l'aide des méthodes de combinatoire analytique, pourquoi cet algorithme de compression de données s'approche, voire atteint, le taux entropique. Les méthodes déployées dans ce chapitre doivent permettre d'étendre le résultat sur le taux de compression de Crochemore *et alii* à toute source sans mémoire. Autre question : quelle est la taille du texte compressé en moyenne et en distribution sur un texte initial de taille n ?

Le résultat final a été obtenu sous un modèle probabiliste sans mémoire de génération des textes. L'analyse du comportement de la taille de l'anti-dictionnaire sous des modèles plus proches de la « réalité » (sources markoviennes, sources dynamiques) est nécessaire à l'analyse des performances de l'algorithme `DCA`. Cette extension présente de nombreux défis, par exemple, le fait que la probabilité du choix des extrémités α et β du motif w dépend désormais des lettres $w_2, \ldots, w_{k-1}$. Le chapitre 6 montre que les méthodes employées dans cette thèse sont aussi valable pour le modèle markovien. Le problème essentiel du cas markovien réside probablement dans le fait que les probabilités des extrémités d'un mot ne sont plus indépendantes du reste du motif.

Un MMI (pour le texte T) est défini comme un mot de taille k n'apparaissant pas dans le texte mais dont les préfixe et suffixe de taille $k-1$ apparaissent. D'autres ensembles de mots minimaux peuvent être envisagés. Il faut garder à l'esprit la difficulté de trouver des anti-dictionnaires de taille optimale : réduire la taille de l'anti-dictionnaire, empêche la suppression de certaines lettres redondantes ; augmenter la taille de l'anti-dictionnaire diminue le taux de compression.

Un autre type d'anti-dictionnaire peut être intéressant à étudier. L'ensemble interdit $D(T)$ pour le texte T est alors formé des mots w n'apparaissant pas dans le texte T pour lesquels le motif $u(w) = w_2 \ldots w_{k-1}$ apparaît à deux reprises (où $w = k$) dans le texte T. La taille moyenne de l'anti-dictionnaire est vraisemblablement asymptotiquement linéaire. La définition de ce nouveau type d'anti-dictionnaire est malléable. On peut demander à ce que les motifs $u(w)$ apparaissent à *au moins* ou à *exactement* deux reprises. Il est aussi possible de discuter du nombre d'occurrences des motifs $u(w)$ dans la définition de l'anti-dictionnaire pour que le taux de compression soit optimal, c'est-à-dire entropique.

Crochemore et Navarro [CN02] ont développé une idée déjà présente dans [CMRS99] et [CMRS00] en introduisant les *presque MMIs* (dans le texte original *almost antifactors*). Leur travail part de la remarque que certains mots apparaissent rarement dans le texte et qu'il est astucieux de considérer certains (choisis selon un critère d'efficacité) de ces mots comme des MMIs et de coder leurs quelques apparitions dans un fichier d'exceptions. Les simulations numériques de [CN02] montrent que l'algorithme améliore le taux de compression ou la place utilisée par rapport à l'algorithme `DCA` « classique ». Combien de mots sont des presque MMIs ? Combien de bits font-ils gagner dans la compression ? Quel est le nombre d'occurrence maximal pour être un presque MMI ?

L'article [CMRS99] mentionne une version dynamique qui construction l'anti-dictionnaire et comprime le texte en parallèle. La définition d'un MMI est uniquement relative à la partie du texte déjà traité. Peut-on mettre en place une version dynamique de la décompression, c'est-à-dire un algorithme dans lequel l'anti-dictionnaire est reconstruit à partir du code en même temps que le texte ? Quelles sont les performances de l'algorithme dynamique par rapport à l'algorithme statique ?

L'algorithme perd de son efficacité quand l'alphabet a une taille supérieure à 2 car on ne peut plus dire « si ce n'est pas la lettre `0`, c'est nécessairement la lettre `1` qui apparaît. » Peut-on proposer des améliorations de l'algorithme dans ce cas ? Peut-on coupler l'algorithme avec un mécanisme de fenêtre coulissante ? Peut-on traiter des morceaux en bloc et accélérer la décompression ?

Chapitre 5

Profondeur typique

Nous montrons que sous un modèle de Markov d'ordre un, la profondeur moyenne d'un arbre des suffixes construit sur les n premiers suffixes se comporte asymptotiquement de manière similaire à la profondeur moyenne d'un trie construit sur n chaînes. Nous établissons ainsi un comportement asymptotique en $(\log n)/h + C$ pour la profondeur moyenne d'un arbre des suffixes, où h est l'entropie du modèle de Markov et C une constante déterminée explicitement. La preuve compare les séries génératrices de la profondeur dans un trie et dans un arbre des suffixes et nous montrons que la différence entre ces deux séries génératrices est asymptotiquement petite. Le résultat de Jacquet and Szpankowski ([JS91]) sur le comportement asymptotique de la profondeur moyenne dans un trie sous modèle de Markov nous permet de conclure.

We prove that under a Markovian model of order one, the average depth of suffix trees of index n is asymptotically similar to the average depth of tries (a.k.a. digital trees) built on n independent strings. This leads to an asymptotic behavior of $(\log n)/h + C$ for the average of the depth of the suffix tree, where h is the entropy of the Markov model and C an explicitly computed constant. Our proof compares the generating functions for the depth in tries and in suffix trees; the difference between these generating functions is shown to be asymptotically small. We conclude using the asymptotic behavior of the average depth in a trie under the Markov model found by Jacquet and Szpankowski ([JS91]).

Sommaire

5.1 Introduction

Le paramètre étudié dans cette section est la *profondeur typique* d'un arbre des suffixes. Le modèle de source qui engendre le texte est plus élaboré et plus puissant que dans les chapitres précédents : il s'agit d'un modèle *markovien*. Nous montrons que le comportement asymptotique de la moyenne de la profondeur typique dans un arbre des suffixes sous modèle de Markov se comporte en $(1/h)\log n + C$. La ligne directrice de la preuve est similaire à celle du chapitre 3 : s'appuyer sur les résultats existants dans le cas d'un trie et estimer le comportement asymptotique de la différence entre la moyenne de la profondeur typique pour un trie et pour un arbre des suffixes. Les techniques utilisées dans ce chapitre sont plus fines que celles du chapitre 3.

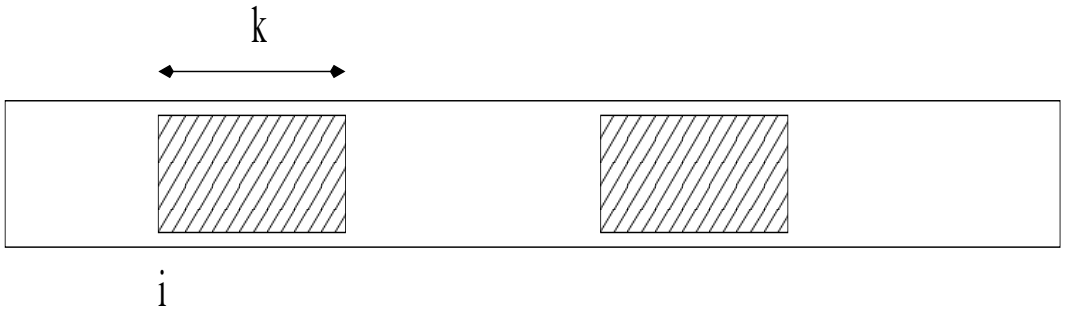


FIG. 5.1 – Recherche du plus long motif commençant à la position i qui se retrouve dans le texte.

Le comportement asymptotique de la moyenne de la longueur de cheminement L_n dans un arbre des suffixe a été obtenu dans le théorème 4 du chapitre 3 sous un modèle de source sans mémoire et avec une forte contrainte sur la probabilité p ($p \in [0.5, 0.54]$). La longueur de cheminement L_n et la profondeur typique D_n d'un arbre des suffixes construit sur n suffixes sont liées par la relation

$$\mathbb{E}(L_n) = n\mathbb{E}(D_n). \tag{5.1}$$

Le résultat du théorème 9 de ce chapitre sur la profondeur typique dans un modèle de Markov d'ordre 1 améliore le résultat du théorème 4 : il étend le domaine de validité du théorème 4 au cas d'une source markovienne et, par conséquent (une source sans mémoire est un sous-cas du modèle markovien), il enlève toute contrainte sur la probabilité dans le cas sans mémoire.

Soit un arbre quelconque avec n feuilles numérotées, la profondeur $D_n(i)$ de la i^{e} feuille est définie comme la longueur du chemin qui va de la racine de l'arbre à la feuille. La **profondeur typique** D_n est la profondeur d'une feuille choisie uniformément au hasard parmi les n feuilles d'un arbre choisi lui aussi au hasard. Le modèle probabiliste est le produit du modèle probabiliste sur le choix de la feuille et du modèle probabiliste sur le choix de l'arbre. Il sera défini de manière plus précise dans la section 5.3. La variable aléatoire D_n nous renseigne sur le profil de l'arbre, alors que la hauteur nous informe sur la feuille la plus profonde.

La moyenne de la profondeur typique d'un arbre des suffixes construit sur les n premiers suffixes d'un texte T est le temps moyen (nombre de lettres lues) nécessaire pour insérer un nouveau suffixe dans l'arbre des suffixes. Lors de l'étape de compression dans l'algorithme LZ'77 (cf. section 1.3), le texte est découpé en *phrases* de telle sorte que la phrase courante soit la plus petite jamais vue dans le texte déjà comprimé. La moyenne de la profondeur typique est la longueur moyenne de ces phrases sur un texte de taille n, c'est-à-dire aussi le nombre moyen de lettres nécessaires pour différencier un suffixe des $n-1$ autres. La profondeur typique sert par exemple à déterminer la redondance dans les codes obtenus par l'algorithme LZ'77.

Dans un arbre des suffixes construit sur un texte T, la profondeur $D_n(i)$ vaut la longueur du plus long facteur commençant à la position i du texte et qui est revu au moins une fois dans le texte (cf. figure 5.1). Soit la plus grande valeur de k pour laquelle il y a $j \neq i$ avec $1 \leq j \leq n$ pour lequel $T_i^{i+k-1} = T_j^{j+k-1}$. Ainsi, $D_n(i) \geq k$ signifie qu'il existe au moins une autre occurrence du motif T_i^{i+k-1} dans le texte T.

Le modèle de source markovienne signifie que la distribution de la lettre aléatoire émise par la source à l'instant t dépend des lettres déjà émises par le passé. Le modèle de source markovienne d'ordre 1 est le plus simple au niveau des calculs. Il signifie que la lettre émise au temps t est une variable aléatoire dont la distribution dépend uniquement de la lettre émise au temps $t-1$.

Les résultats que nous montrons sous modèle markovien d'ordre 1 se généralisent au modèle de sources markoviennes d'ordres supérieurs, sans introduire de difficultés conceptuelles nouvelles à l'exception d'une saturation d'indices dans les notations.

Nous considèrons que le lecteur connait les définitions de la matrice de transition, du vecteur stationnaire et autres définitions de base des chaînes de Markov (des milliers d'ouvrages le font très bien, la référence la plus classique est le livre de William Feller [Fel68]). Les textes que l'on regarde sont suffisamment longs pour que la source soit supposée *stationnaire*. Le vecteur des probabilités stationnaires est noté $\pi = (\pi_i)_i$. La matrice de transition qui code la probabilité d'avoir une lettre donnée au temps t sachant la lettre précédente est notée $P = (p_{i,j})_{i,j}$. La probabilité p_w est (toujours) la probabilité que le texte émis par la source commence par le mot w. La matrice stationnaire de probabilité est notée Π, chacune de ses colonnes est le vecteur de probabilité stationnaire π. La chaîne de Markov sous-jacente est supposée *irréductible* et *apériodique* (cette notion d'apériodique n'a rien à voir ni avec la périodicité des pôles de la transformée de Mellin, ni avec la périodicité d'un motif au sens de Guibas et Odlyzko [GO81a]).

Ce chapitre compare les profondeurs typiques d'un arbre des suffixes construit sur les n premiers suffixes d'un texte T et d'un trie construit sur n chaînes indépendantes. On introduit les notations D_n pour la profondeur typique d'un arbre des suffixes et D_n^t pour celle d'un trie.

Ce chapitre présente une nouvelle difficulté par rapport aux chapitres précédents : en plus de la dépendance combinatoire entre les suffixes d'un texte (la corrélation), il existe dans la probabilité même du suffixe une dépendance en fonction des symboles précédents (modèle de Markov).

La moyenne de la profondeur typique d'une feuille d'un trie sous modèle de Bernoulli est déjà dans Knuth [Knu73]. Devroye étend le résultat à un modèle à densité beaucoup plus puissant dans [Dev82]. Flajolet et Sedgewick [FS86] retrouvent le même résultat que Knuth en utilisant la méthode de Rice. Pittel [Pit85] développe des résultats sur la convergence en probabilité et oscillation presque-sûre de la profondeur typique dans un trie sous un modèle très général. Dans [Pit86] il obtient la distribution limite de la profondeur typique. La variance a été obtenue par Kirschenhofer et Prodinger [KP88] dans le cas symétrique binaire. Szpankowski [Szp88] étend le résultat sur la variance au cas d'alphabets de taille finie. En 1991, Jacquet et Szpankowski [JS91] obtiennent le comportement asymptotique de la profondeur moyenne d'un trie sous un modèle de source markovienne d'ordre 1.

Théorème [Jacquet et Szpankowski '91] *Dans une source stationnaire markovienne d'ordre 1, la moyenne de la profondeur typique se comporte asymptotiquement en*

$$\mathbb{E}(D_n^t) = \frac{1}{h}\left(\log n + \gamma + \frac{h_2}{2h} - H + \epsilon(n)\right) + O(1/n),$$

$$\text{où} \qquad h := -\sum_{(i,j)\in\mathcal{A}^2} \pi_i p_{i,j} \log p_{i,j}, \quad H := -\sum_{i\in\mathcal{A}} \pi_i \log \pi_i \ \text{ et}$$

$$h_2 := h_2(1) \ \text{ avec } h_2(s) := \lim_{k\to\infty} \frac{d^2}{ds^2}\left(\sum_{w\in\mathcal{A}^k} p_w^s\right)^{1/k}.$$

Devroye [Dev92] obtient (entre autres) la loi limite de la profondeur typique dans un trie sous un modèle à densité. Dans [Dev02], il trouve des inégalités de concentration sur les paramètres du trie (dont la profondeur typique).

Wyner et Ziv [WZ89] montrent que la profondeur typique (pour une source stationnaire et ergodique) d'un arbre des suffixes converge en probabilité vers $(1/h)\log n$. Ils conjecturent que la convergence de la profondeur typique est même valable presque sûrement. Dans [Szp93a], Szpankowski invalide leur conjecture : la profondeur typique oscille presque sûrement et dans [Szp93b], il montre la convergence en probabilité de la profondeur typique pour un arbre des suffixes généralisé. Un comportement asymptotique plus précis de la moyenne de la profondeur typique (avec aussi d'autres paramètres) dans un arbre des suffixes sous un modèle de source sans mémoire est obtenu dans Jacquet et Szpankowski [JS94]. Ils utilisent une méthode faisant appel aux alignements. On y trouve aussi la variance et la distribution limite de la profondeur typique dans un arbre des suffixes. Wyner [Wyn97] trouve le terme dominant du comportement asymptotique de la moyenne et la distribution de la profondeur typique dans la version FDLZ (*Fixed-Database Lempel-Ziv*, le dictionnaire est fixé au départ et n'évolue pas) de l'algorithme LZ'77. Cette version évite tous les problèmes de corrélation qui sont au cœur de l'analyse de l'algorithme de Lempel et Ziv. Ses résultats sont valables pour des sources markoviennes. Enfin Jacquet et Szpankowski [JS05] retrouvent par des méthodes analytiques le comportement asymptotique de la profondeur typique dans un arbre des suffixes sous un modèle de source sans mémoire.

Ce chapitre se focalise sur la longueur des phrases dans le découpage d'un texte par l'algorithme LZ'77. L'algorithme-frère de compression de données sans pertes LZ'78 se base sur les arbres digitaux de recherche ou *digital search trees* ou DST. Les paramètres des DST ont déjà fait l'objet de plusieurs études [GK92, JS95]. Louchard et Szpankowski [LS95] obtiennent la moyenne de la profondeur typique, sa variance et sa distribution pour un modèle de source sans mémoire. Louchard, Szpankowski et Tang [LST99] réalisent l'étude sous un modèle sans mémoire de la profondeur typique dans un DST généralisé. Jacquet, Szpankowski et Tang [JST01] trouvent la moyenne $\mathbb{E}^{\mathrm{DST}}(D)$, la variance et la distribution limite de la profondeur typique dans un DST construit sur n phrases sous modèle markovien :

$$\mathbb{E}^{\mathrm{DST}}(D) = \frac{\log n}{h} + O(1).$$

On souhaite valider la même intuition que dans le cas de l'étude de la taille et de la longueur de cheminement, à savoir que la profondeur moyenne d'un trie construit sur n mots engendrés par une source markovienne d'ordre 1 et d'un arbre des suffixes construit sur les n premiers suffixes d'un texte engendré par la même source se comportent de manière similaire asymptotiquement. D'abord une expression asymptotique de la différence entre profondeur typique d'un arbre des suffixes et d'un trie sous modèle de Markov d'ordre 1 est obtenue. Ensuite, le comportement asymptotique de cette différence est déterminé. Le résultat de [JS91] sur le comportement asymptotique de la moyenne de la profondeur typique d'un trie sous modèle de Markov permet de conclure que pour un arbre des suffixes construit sur les n premiers suffixes d'un texte produit par une source markovienne d'ordre 1, le comportement asymptotique de la moyenne de la profondeur typique est de l'ordre de $(\log n)/h + C$. Le schéma de la preuve suit celui débroussaillé dans [JS05] pour la profondeur typique d'un arbre des suffixes dans un modèle de source sans mémoire.

Dans la section 5.2, nous montrons que le polynôme d'auto-corrélation $c_w(z)$ du motif w vaut presque 1 avec haute probabilité. Notre but étant de comparer la profondeur typique dans un arbre des suffixes et dans un trie, on introduit les séries génératrices bivariées $D(z,u)$ de D_n et $D^t(z,u)$ de D_n^t dans la section 5.3. Nous avons

$$D(z,u) = \frac{1-u}{u} \sum_{w\in\mathcal{A}^\star} (zu)^{|w|} \frac{p_w}{\mathfrak{D}_w(z)^2} \qquad \text{et} \qquad D^t(z,u) = \frac{1-u}{u} \sum_{w\in\mathcal{A}^\star} u^{|w|} \frac{zp_w}{(1-z+zp_w)^2},$$

où $\mathfrak{D}(z)$ est une série génératrice prenant en compte le caractère markovien de la source, introduite dans [RS98].

La section 5.3 permet de voir que $D^t(z,u)$ a exactement un pôle d'ordre 2 pour chaque motif w. Une fois certaines conditions satisfaites, le théorème de Rouché appliqué dans la section 5.4.1 affirme que pour z dans un disque de rayon $\rho > 1$ et pour des motifs de taille suffisamment grande, la série génératrice $\mathfrak{D}_w(z)$ a exactement une racine dominante d'ordre 2 pour chaque w. Ainsi $D(z,u)$ a exactement un pôle dominant d'ordre 2 pour chaque w de taille suffisamment grande. La différence $D(z,u) - D^t(z,u)$ entre les deux séries génératrices bivariées est analysée dans la section 5.4.2 à l'aide de méthodes d'analyse complexe déjà rencontrées. Le théorème de Cauchy permet de déterminer la contribution de chacun des pôles à la différence

$$Q_n(u) := u(1-u)^{-1}[z^n](D(z,u) - D^t(z,u)).$$

L'analyse asymptotique de la différence $Q_n(u)$ est effectuée à l'aide de la théorie de Mellin.

À partir de ces résultats, nous pouvons conclure que les moyennes des profondeurs typiques D_n et D_n^t sont asymptotiquement similaires. Ainsi, D_n a une moyenne de l'ordre de $\log n/h$ avec une très faible fluctuation périodique. Plus précisément,

Théorème 9 *Pour un arbre des suffixes construit sur les n premiers suffixes d'un texte produit par une source de Markov d'ordre 1, la moyenne de la profondeur typique se comporte asymptotiquement en*

$$\mathbb{E}_n(D_n) = \frac{1}{h}\left(\log n + \gamma + \frac{h_2}{2h} - H + \epsilon(n)\right) + O(n^{-c}),$$

où c est un réel positif dépendant du modèle probabiliste, γ la constante d'Euler, h l'entropie de la source de Markov, h_2 la deuxième entropie, H l'entropie stationnaire et $\epsilon(z)$ une fonction de très faible amplitude fluctuant autour de zéro. En particulier

$$h := -\sum_{(i,j)\in\mathcal{A}^2} \pi_i p_{i,j} \log p_{i,j}, \quad H := -\sum_{i\in\mathcal{A}} \pi_i \log \pi_i \ \text{ et}$$
$$h_2 := h_2(1) \ \text{ avec } h_2(s) := \lim_{k\to\infty} \frac{d^2}{ds^2}\left(\sum_{w\in\mathcal{A}^k} p_w^s\right)^{1/k}. \tag{5.2}$$

5.2 Auto-corrélation

Dans cette section, nous allons montrer quelques propriétés supplémentaires sur l'auto-corrélation qui nous seront utiles dans ce chapitre. D'abord nous raffinons le lemme 11 de la section 2.4.3 qui donnait une borne sur le nombre de motifs dont la valeur en 1 du polynôme d'auto-corrélation était « loin » de 1. Ensuite, le résultat du lemme 8 de la section 2.3.2 page 45 qui prouvait que le polynôme d'auto-corrélation ne s'annulait pas dans un disque centré de rayon plus grand que 1 est étendu au cas d'une source de Markov d'ordre 1.

Soit w le facteur du texte T qui coïncide le plus longtemps avec le suffixe commençant à la position i et qui se retrouve dans le texte ailleurs qu'à la position i. Il est possible que deux occurrences de ce plus long motif w se chevauchent. Cette possibilité crée évidemment des complications dans la détermination de la probabilité des événements $\{D_n(i) \geq k\}$. L'auto-corrélation joue alors un rôle central. Le polynôme d'auto-corrélation probabilisé $c_w(z)$ de la

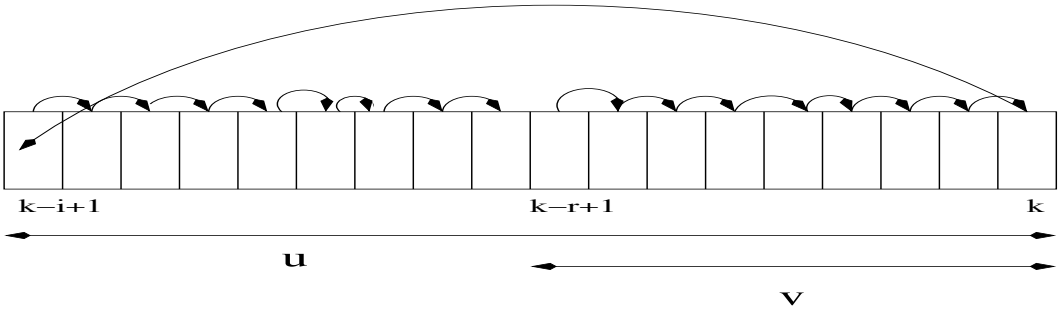


FIG. 5.2 – Le suffixe u et les probabilités de transitions utiliséees dans l'équation (5.5).

chaîne w a été défini dans le cas d'une source sans mémoire par l'équation (2.4) de la page 41. Dans le cas d'une source markovienne, il est défini par

$$c_w(z) := \sum_{j=0}^{k-1} c_j \mathbb{P}(w_{k-j+1}^k | w_{k-j}) z^j. \tag{5.3}$$

On vérifie aisément que les deux définitions coïncident dans un modèle de source sans mémoire.

Une *période* d'un motif w est un entier i pour lequel $c_i = 1$ (un motif a généralement plusieurs périodes). La période minimale $m(w)$ de w est la plus petite période strictement positive. Un motif a toujours une période nulle ($c_0 = 1$). Néanmoins il se peut qu'il n'y ait pas de période strictement positive pour un motif w. On fixe alors par convention $m(w) = k$. Un motif est considéré comme peu corrélé si sa période minimale est grande.

Le lemme suivant formule de manière précise l'intuition que pour la plupart des motifs d'une taille k, la valeur en 1 du polynôme d'auto-corrélation probabilisé est très proche de 1. Cela résulte du fait que la somme sur les motifs peu corrélés des probabilités p_w est très proche de 1.

Lemme 41 *Il existe $\theta > 0$, $\delta < 1$ et pour tout $\rho > 1$ vérifiant $\rho\delta < 1$ et $p\rho < 1$, on a pour tout entier k*

$$\sum_{w \in \mathcal{A}^k} [\![|c_w(\rho) - 1| \leq (\rho\delta)^k \theta]\!] p_w \geq 1 - \theta\delta^k. \tag{5.4}$$

Preuve : Notons que le polynôme $c_w(z) - 1$ a un terme de degré i avec $1 \leq i \leq k-1$ si et seulement si $c_i = 1$. Si $c_i = 1$, le suffixe de taille i de w se répète dans w autant de fois qu'il y a de i dans k (tout ceci a déjà été développé dans la preuve du lemme 8 de la section 2.3.2 et les figures 2.3 et 2.2 adjacentes, page 45). Sachant que $c_i = 1$, la connaissance des i dernières lettres de w permet de déterminer complètement le mot w. Ainsi, pour un suffixe $w_{k-i+1} \dots w_k$, il existe un unique mot $w_1 \dots w_{k-i}$ tel que le polynôme $c_w(z) - 1$ ait degré $i \leq k-1$.

On définit $\tilde{p}_{i,j} := p_{w_i,w_j}$, $\tilde{\pi}_i := \pi_{w_i}$ et $p := \max_{1 \leq i,j \leq 2} (\tilde{p}_{i,j}, \tilde{\pi}_i)$. En s'aidant de la figure 5.2 et pour j et k fixés, il vient

$$\begin{aligned}
\sum_{i=1}^{j} \sum_{w \in \mathcal{A}^k} [\![m(w) = i]\!] p_w &= \sum_{i=1}^{j} \sum_{w_1,\dots,w_k \in \mathcal{A}^k} [\![m(w) = i]\!] \tilde{\pi}_1 \tilde{p}_{1,2} \cdots \tilde{p}_{k-i-1,k-i} \tilde{p}_{k-i,k-i+1} \cdots \tilde{p}_{k-1,k} \\
&\leq \sum_{i=1}^{j} \sum_{w_{k-i+1},\dots,w_k \in \mathcal{A}^i} \tilde{\pi}_1 \tilde{p}_{k-i,k-i+1} \cdots \tilde{p}_{k-r-1,k-r} \tilde{p}_{k-r+1,k-r+2} \cdots \tilde{p}_{k-1,k} p^{k-i}.
\end{aligned} \tag{5.5}$$

Si $m(w) = i$, le suffixe de taille i détermine complètement le mot w. Donc la somme porte en fait sur les i dernières lettres $w_{k-i+1}, \cdots, w_k$. Le produit des probabilités $\tilde{p}_{1,2} \cdots \tilde{p}_{k-i-1,k-i}\tilde{p}_{k-r,k-r+1}$ est trivialement majoré par p^{k-i}. Le suffixe u se répète dans w et on sait que w_1 est la première lettre du motif v (suffixe de u de taille $r := k - \lfloor k/i \rfloor i$). On récrit le terme général de la somme :

$$\begin{aligned}\tilde{\pi}_1\tilde{p}_{k-i+1,k-i+2}\cdots\tilde{p}_{k-1,k}p^{k-i} &= \pi_{k-r+1}\tilde{p}_{k-r+1,k-r+2}\cdots\tilde{p}_{k-1,k}\tilde{p}_{k-i,k-i+1}\cdots\tilde{p}_{k-r-1,k-r}p^{k-i}\\ &= \pi_{k-r+1}\tilde{p}_{k-r+1,k-r+2}\cdots\tilde{p}_{k-1,k}\tilde{p}_{k,k-i+1}\cdots\tilde{p}_{k-r-1,k-r}p^{k-i},\end{aligned}$$

puisque par répétition du motif u, $w_{k-i} = w_k$ et $w_{k-i+1} = w_1$. Ensuite, on somme sur toutes les lettres du suffixe u

$$\sum_{w_{k-i+1},\ldots,w_k \in \mathcal{A}^i} \tilde{\pi}_{k-r+1}\tilde{p}_{k-r+1,k-r+2}\cdots\tilde{p}_{k-1,k}\tilde{p}_{k,k-i+1}\cdots\tilde{p}_{k-r-1,k-r} = \Pi P^{i-1}\mathbf{1} = 1. \tag{5.6}$$

À partir de ce résultat, on obtient

$$\sum_{w\in\mathcal{A}^k} [\![c_w(z) - 1 \text{ a degré minimal } \leq j]\!]p_w = \sum_{w\in\mathcal{A}^k}\sum_{i=1}^{j}[\![m(w) = i]\!]p_w \leq \sum_{i=1}^{j} p^{k-i} \leq \frac{p^{k-j}}{1-p}, \tag{5.7}$$

et ceci reste vrai pour $j = \lfloor k/2 \rfloor$. Ainsi

$$\sum_{w\in\mathcal{A}^k} [\![\text{tous les termes de } c_w(z) - 1 \text{ ont degré} > \lfloor k/2 \rfloor]\!]p_w \geq 1 - \frac{p^{\lceil k/2 \rceil}}{1-p} = 1 - \theta\delta^k. \tag{5.8}$$

Si tous les termes de $c_w(z) - 1$ ont un degré supérieur à $\lfloor k/2 \rfloor$, alors pour tout ρ tel que $\rho\delta < 1$, on a

$$|c_w(\rho) - 1| \leq \sum_{i=\lfloor k/2 \rfloor}^{k} (\rho p)^i \leq \rho^k \frac{p^{\lfloor k/2 \rfloor}}{1-p} = (\rho\delta)^k\theta. \tag{5.9}$$

Le choix de $\delta = \sqrt{p}$, $\theta = (1-p)^{-1}$ et un $\rho > 1$ qui vérifie $\delta\rho < 1$ et $p\rho < 1$ permet de conclure la preuve du lemme. ◄

Le lemme suivant montre que dans un disque centré de rayon $\phi > 1$ et pour des motifs w de taille suffisamment grande, le polynôme d'auto-corrélation $c_w(z)$ ne s'annule pas. Ce lemme est une extension du lemme 8 au cas de sources markoviennes d'ordre 1.

Lemme 42 *Pour tout $\phi > 1$ tel que $p\phi < 1$, il existe un entier K et $\alpha > 0$ tels que, pour chaque motif w de taille plus grande que K et pour z dans le disque de rayon ϕ, on ait*

$$|c(z)| \geq \alpha.$$

Preuve : Comme pour le lemme précédent, on divise la preuve du lemme en deux selon la valeur de la période minimale $m(w)$ du motif w de taille k. Soit un réel $\phi > 1$ qui vérifie $p\phi < 1$ avec $p = \max\{p_{i,j}, \pi_i\}$. Puisque le coefficient c_0 du polynôme d'auto-corrélation vaut toujours 1, le polynôme d'auto-corrélation s'écrit

$$c_w(z) = 1 + \sum_{j=m(w)}^{k-1} c_j \mathbb{P}(w_{k-j+1}^k | w_{k-j}) z^j. \tag{5.10}$$

Si $m(w) > \lfloor k/2 \rfloor$ alors

$$|c_w(z)| \geq 1 - \left| \sum_{j=m(w)}^{k-1} c_j \mathbb{P}(w_{k-j+1}^k | w_{k-j}) z^j \right| \geq 1 - \frac{(p\phi)^{m(w)}}{1-p\phi}, \tag{5.11}$$

dans le disque $|z| \leq \phi$. Pour plus de facilité, on note $l = m(w)$. Puisque $l > \lfloor k/2 \rfloor$ et $p\phi < 1$, le polynôme d'auto-corrélation est minoré par

$$|c_w(z)| \geq 1 - \frac{(p\phi)^{k/2}}{1-p\phi}. \tag{5.12}$$

Pour des motifs de taille K_1 suffisamment grande, la quantité $(p\phi)^{k/2}$ est extrêmement faible et le polynôme d'auto-corrélation probabilisé est supérieur à une constante strictement positive α.

Rappelons (cf. la preuve du lemme 8, chapitre 3 et les figures 2.2 et 2.3 adjacentes) que si $c_l = 1$, le suffixe $u = w_{k-l+1} \cdots w_k$ de taille l de w va se répeter dans w autant de fois qu'il y a l dans k, c'est-à-dire $q := \lfloor k/l \rfloor$ fois. Le reste, si le motif u ne peut se répéter un nombre entier de fois, est le suffixe v de u de taille $r := k - \lfloor k/l \rfloor l$, et alors $w = vu^{\lfloor k/l \rfloor}$.

Si $m(w) = l \leq \lfloor k/2 \rfloor$, le polynôme d'auto-corrélation s'écrit explicitement en tenant compte des répétitions :

$$c_w(z) = 1 + \mathbb{P}(u|w_{k-l}) z^l + \mathbb{P}(uu|w_{k-l}) z^{2l} + \cdots + \mathbb{P}(u^{q-1}|w_{k-l}) z^{l(q-1)} c_{vu}(z). \tag{5.13}$$

Il est plus difficile d'avoir une expression des derniers termes dans le polynôme d'auto-corrélation $c_w(z)$, on laisse donc le polynôme d'auto-corrélation $c_{vu}(z)$ sans le développer. Nous écrivons la probabilité $\mathbb{P}(u|w_{k-l})$ comme le produit des l probabilités de transition

$$A := \mathbb{P}(u|w_{k-l}) = p_{w_{k-l},w_{k-l+1}} p_{w_{k-l+1},w_{k-l+2}} \cdots p_{w_{k-1},w_k},$$

d'où

$$\begin{aligned} \mathbb{P}(u^j|w_{k-l}) &= p_{w_{k-l},w_{k-l+1}} p_{w_{k-l+1},w_{k-l+2}} \cdots p_{w_{k-1},w_k} (p_{w_{k-l},w_{k-l+1}} \cdots p_{w_{k-1},w_k})^{j-1} \\ &= A^j. \end{aligned}$$

Nous obtenons l'écriture du polynôme d'auto-corrélation

$$c_w(z) = 1 + Az^l + \left(Az^l\right)^2 + \cdots + \left(Az^l\right)^{q-1} c_{vu}(z) = \frac{1 - \left(Az^l\right)^{q-1}}{1 - Az^l} + \left(Az^l\right)^{q-1} c_{vu}(z). \tag{5.14}$$

Cette expression permet de trouver une borne inférieure pour $|c_w(z)|$:

$$\begin{aligned} |c_w(z)| &\geq \left| \frac{1 - \left(Az^l\right)^{q-1}}{1 - Az^l} \right| - \left| \left(Az^l\right)^{q-1} c_{vu}(z) \right| \geq \frac{1 - (p\phi)^{l(q-1)}}{1 + (p\phi)^l} - (p\phi)^{l(q-1)} |c_{vu}(z)| \\ &\geq \frac{1 - (p\phi)^{l(q-1)}}{1 + (p\phi)^l} - \frac{(p\phi)^{l(q-1)}}{1 - p\phi}. \end{aligned}$$

Par définition de ϕ, le produit $p\phi$ est inférieur à 1 ainsi $(p\phi)^k$ tend vers zéro quand k tend vers l'infini. Le produit $l(q-1)$ est proche de k (au pire il vaut $k/3$ si le motif se décompose en $w = uuv$). Donc $(p\phi)^k$ tend vers zéro et pour des motifs w de taille supérieure à un entier K_2 suffisamment grand, seul le terme $(1 + (p\phi)^l)^{-1}$ reste dans minoration. Il existe donc un réel positif α tel que le polynôme d'auto-corrélation soit supérieur à α. Le résultat final est obtenu en posant $K = \max\{K_1, K_2\}$. ◄

5.3 Les séries génératrices

Le modèle probabiliste sur la variable aléatoire D_n (resp. D_n^t), profondeur typique d'un arbre des suffixes construit sur les n premiers suffixes d'un texte (resp. trie construit sur n textes indépendants), est le produit

- d'un modèle de Markov d'ordre 1 pour la source qui engendre le texte (resp. les n textes) et ;
- d'un modèle uniforme sur $\{1, \ldots, n\}$ pour le choix d'une feuille dans l'arbre des suffixes (resp. trie).

Si X est une variable aléatoire distribuée uniformément sur $\{1, \ldots, n\}$, $\mathcal{T}$ est l'arbre des suffixes construit sur les n premiers suffixes d'un texte aléatoire T engendré par la source et $\mathcal{T}^t$ est le trie construit sur n textes aléatoires indépendants engendrés par la même source, alors la profondeur typique dans un arbre des suffixes et dans un trie sont définies par

$$D_n := \sum_{i=1}^{n} D_n(i)(\mathcal{T})[\![X = i]\!] \quad \text{et} \quad D_n^t := \sum_{i=1}^{n} D_n^t(i)(\mathcal{T}^t)[\![X = i]\!]. \tag{5.15}$$

L'exposant t sur une quantité indique qu'il s'agit d'une quantité relative au trie.

Notre objectif est de comparer asymptotiquement les séries génératrices probabilisées de la profondeur pour un arbre des suffixes et pour un trie. La série génératrice de la profondeur pour un arbre des suffixes est définie par

$$D_n(u) := \sum_k \mathbb{P}(D_n = k)u^k, \tag{5.16}$$

et celle pour un trie par

$$D_n^t(u) := \sum_k \mathbb{P}(D_n^t = k)u^k. \tag{5.17}$$

Nous obtenons dans les paragraphes suivants une expression explicite de ces séries génératrices et de leur séries génératrices bivariées (où z compte la longueur du texte) associées

$$D(z, u) := \sum_n nD_n(u)z^n \text{ et } D^t(z, u) := \sum_n nD_n^t(u)z^n. \tag{5.18}$$

➜ Les séries génératrices pour les tries

Nous trouvons d'abord l'expression de la série génératrice $D_n^t(u)$. Chaque chaîne de l'ensemble de base $S = \{T_1, \cdots, T_n\}$ est associée à une unique feuille dans le trie. Par définition d'un trie, les lettres lues sur le chemin allant de la racine du trie à la feuille qui contient la chaîne forment le plus petit préfixe qui permet de distinguer la chaîne en question des $n - 1$ autres. On choisit aléatoirement une feuille i dans le trie. Il existe une unique chaîne T_i dont le préfixe amène à cette feuille. Ce préfixe de taille k de T_i est noté w. La variable aléatoire $D_n^t(i)$ est inférieure à k si et seulement si aucune des $n - 1$ chaînes de l'ensemble S autre que T_i ne commence par le préfixe w. Ainsi, on écrit

$$\mathbb{P}(D_n^t(i) < k) = \sum_{w \in \mathcal{A}^k} p_w(1 - p_w)^{n-1}.$$

La série génératrice probabilisée $D_n^t(u)$ vaut

$$\begin{aligned} D_n^t(u) &= \sum_{k\geq 0} \mathbb{P}(D_n = k)u^k = \sum_{k\geq 0} \left(\mathbb{P}(D_n < k+1) - \mathbb{P}(D_n < k)\right) u^k \\ &= \sum_{k\geq 0} \mathbb{P}(D_n < k+1)u^k - \sum_{k\geq 0} \mathbb{P}(D_n < k)u^k \\ &= \sum_{k\geq 0} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(D_n^t(i) < k+1)u^k - \sum_{k\geq 0} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(D_n^t(i) < k)u^k \\ &= \frac{1-u}{u} \sum_{w\in\mathcal{A}^\star} u^{|w|} p_w (1-p_w)^{n-1}. \end{aligned} \tag{5.19}$$

L'expression de la série génératrice bivariée $D^t(z,u)$ s'obtient facilement :

$$D^t(z,u) = \frac{1-u}{u} \sum_{w\in\mathcal{A}^\star} u^{|w|} \frac{zp_w}{(1-z+zp_w)^2}, \tag{5.20}$$

pour $|u| < 1$ et $|z| < 1$.

➜ La série génératrice bivariée pour les arbres des suffixes

La série génératrice bivariée pour la profondeur d'un arbre des suffixes (dans le cadre d'une source markovienne) a été obtenue par Jacquet et Szpankowski dans [JS05]. La norme $\|.\|$ désigne une norme quelconque sur les matrices.

Lemme 43 *Pour $|u| < 1$ and $|z| < 1$,*

$$D(z,u) = \frac{1-u}{u} \sum_{w\in\mathcal{A}^\star} (zu)^{|w|} \frac{p_w}{\mathfrak{D}_w(z)^2}, \tag{5.21}$$

où $\mathfrak{D}_w(z) = (1-z)c_w(z) + z^k p_w(1+(1-z)F(z))$, $|w| = k$ et pour $|z| < \|P-\Pi\|^{-1}$,

$$F(z) := \frac{1}{\pi_{w_1}} \left[\sum_{n\geq 0} (P-\Pi)^{n+1} z^n\right]_{w_k,w_1} = \frac{1}{\pi_{w_1}} [(P-\Pi)(I-(P-\Pi)z)^{-1}]_{w_k,w_1}. \tag{5.22}$$

5.4 Asymptotique de la profondeur moyenne

Dans cette section, nous obtenons le comportement asymptotique de la moyenne de la profondeur dans un arbre des suffixes construit sur les n premiers suffixes d'un texte engendré par une source markovienne d'ordre 1. Nous procédons, comme dans la section 2.3 du chapitre 3, à l'aide des techniques de la combinatoire analytique. Dans un premier temps, nous montrons que, pour chaque motif w suffisamment grand, il existe un unique zéro dominant de $\mathfrak{D}_w(z)$ et qu'il est réel et positif. La méthode est identique à celle utilisée dans la section 2.3.2 du chapitre 3, sauf qu'ici la source est markovienne (d'ordre 1). Dans la section 5.4.2, l'application du théorème de Cauchy permet d'obtenir l'expression asymptotique de la différence entre la série génératrice $D_n(u)$ et la série génératrice $D_n^t(u)$ sous forme d'une somme $Q_n(u)$. Le comportement asymptotique de cette somme $Q_n(u)$ est determiné dans la section 5.4.3 à l'aide de la théorie de Mellin.

5.4.1 Localisation du pôle dominant

Nous montrons que pour un motif w de taille suffisamment grande, $\mathfrak{D}_w(z)$ a une unique racine dominante et qu'elle est réelle, positive et d'ordre un. Nous prouvons aussi qu'il existe un disque centré en l'origine et de rayon ϕ plus grand que 1 qui contient, pour chaque motif w de taille suffisamment grande, l'unique zéro dominant de $\mathfrak{D}_w(z)$ et aucun autre zéro de $\mathfrak{D}_w(z)$.

Le théorème de Pringsheim permet de montrer qu'il existe un zéro dominant réel et positif pour $\mathfrak{D}_w(z)$ (cf. la section 2.3.2). Ce zéro est noté A_w par la suite. Le lemme suivant est la généralisation du lemme 7 de la section 2.3.2 page 45 au cas des sources markoviennes (d'ordre 1) :

Lemme 44 *Il existe $\phi > 1$ et un entier K, tel que dans le disque centré de rayon ϕ et pour chaque motif de taille supérieure à K, la fonction $\mathfrak{D}_w(z) = p_w z^k(1 + (1-z)F(z)) + (1-z)c(z)$ ne s'annule qu'une seule fois.*

Preuve : Soit w un motif de taille k. L'énoncé du théorème de Rouché a été rappelé dans la section 2.3.2 du chapitre 3. Les fonctions sur lesquelles on applique le théorème de Rouché sont

$$f(z) := (1-z)c_w(z) \text{ et } g(z) := p_w z^k(1 + (1-z)F(z)).$$

Pour pouvoir appliquer le théorème sur le cercle centré en l'origine de rayon ϕ, il faut que sur ce, les fonctions f et g vérifient $|f(z)| > |g(z)|$. Sur le cercle ces deux fonctions sont analytiques : f est un polynôme et la fonction F qui prend en compte la « markoviannité » de la source est analytique pour $|z| < \left\|P - \Pi\right\|^{-1}$ (où $\left\|P - \Pi\right\|^{-1} > 1$). Donc g est analytique pour $\phi < \left\|P - \Pi\right\|^{-1}$ et cette hypothèse du théorème de Rouché est vérifiée.

La quantité $p_w z^k$ décroît exponentiellement pour des motifs de taille suffisamment grande. Le problème restant est de majorer la fonction $F(z)$ sur le cercle de rayon ϕ. On note $d = \min_{a\in\mathcal{A}} \pi_a$. Cette valeur est nécessairement strictement positive (sinon une lettre n'apparaîtrait jamais). De plus $(P-\Pi)^{n+1} = P^{n+1} - \Pi$ (par définition $P\Pi = \Pi P = \Pi$ et $\Pi\Pi = \Pi$) donc nous obtenons :

$$\begin{aligned} |F(z)| &\leq \frac{1}{d}\left|\left[\sum_{n\geq 0}(P^{n+1} - \Pi)z^n\right]_{w_k,w_1}\right| \leq \frac{1}{d}\sum_{n\geq 0}|[P^{n+1}]_{w_k,w_1} - [\Pi]_{w_k,w_1}||z|^n \\ &\leq \frac{1}{d}\sum_{n\geq 0} b r^{n+1}\phi^n \leq \frac{br}{d}\frac{1}{1-r\phi}, \end{aligned} \tag{5.23}$$

où b et r sont des constantes (indépendantes du motif w) avec $0 < r < 1$ et $|z| = \phi$ tel que $r\phi < 1$ (cf. Kemeny et Snell [KS60]).

Soit K un entier , $\alpha > 0$ et $\rho > 1$ un rayon qui satisfont le lemme 42. On prend en plus $\phi = \min\{\rho, \phi\}$ pour que les deux majorations (majoration de $F(z)$ et minoration de $c_w(z)$) soient valides sur le cercle de rayon ϕ. Puisque ρ vérifie $p\rho < 1$, ϕ aussi. Ainsi, il existe un entier $K' > K$ pour lequel, sur le disque de rayon ϕ, tous les motifs de taille supérieure à K' vérifient

$$|g_w(z)| \leq (p\phi)^k\left(1 + (1+\phi)\frac{br}{d}\frac{1}{1-r\phi}\right) < \alpha(\phi - 1) \leq |f_w(z)|. \tag{5.24}$$

Les hypothèses du théorème de Rouché sont vérifiées pour les fonctions f_w et g_w sur le cercle centré de rayon ϕ. Par conséquent, $(f+g)(z) = \mathfrak{D}_w(z)$ a autant de zéro dans le disque centré de rayon ϕ que $f(z)$, c'est-à-dire un seul puisque $c_w(z)$ ne s'annule pas dans le disque (par construction de ρ et donc de ϕ).

Les hypothèses du théorème de Rouché sont valables sur le cercle de rayon ϕ pour n'importe quel motif de taille supérieure à K'. Donc le disque de rayon ϕ contient, pour chaque motif w de taille supérieure à K', l'unique pôle dominant de $\mathfrak{D}_w(z)$ et aucun autre pôle de $\mathfrak{D}_w(z)$. ◄

5.4.2 Expression de la différence

Cette section a deux objectifs : d'abord obtenir une valeur utilisable du pôle dominant A_w de la fonction $\mathfrak{D}_w(z)$ et ensuite exprimer asymptotiquement, en utilisant le théorème de Cauchy, la différence entre les coefficients des séries génératrices bivariées de la profondeur pour un trie et pour un arbre des suffixes.

Les dérivées première et deuxième de $\mathfrak{D}_w(z)$ en son pôle dominant $z = A_w$ sont appelées B_w et E_w (C_w nous a déjà servi en tant qu'ensemble d'auto-corrélation et D_w risquant de confondre le lecteur étant donné le nombre important de « D » dans cette section, la dérivée deuxième sera notée E_w). La technique de *bootstrapping* déjà utilisée pour approcher le pôle A_w (qui était alors nommé ρ) dans la section 2.3.1 est réutilisée et permet d'obtenir

$$\begin{aligned} A_w &= 1 + \frac{p_w}{c_w(1)} + O\left(p_w^2\right), \\ B_w &= -c_w(1) + p_w\left[k - F(1) - 2\frac{c'_w(1)}{c_w(1)}\right] + O\left(p_w^2\right) \text{ et} \\ E_w &= -2c'_w(1) + p_w\left[-3\frac{c''_w(1)}{c_w(1)} + k(k-1) - 2F'(1) - 2kF(1)\right] + O\left(p_w^2\right). \end{aligned} \tag{5.25}$$

Notre objectif principal est de comparer les séries génératrices $D_n(u)$ et $D_n^t(u)$ de la profondeur dans un arbre des suffixes et dans un trie pour montrer que le comportement asymptotique de la moyenne de la profondeur est similaire dans un trie et dans un arbre des suffixes. La série génératrice au cœur du problème est

$$Q_n(u) := \frac{u}{1-u}(D_n(u) - D_n^t(u)), \tag{5.26}$$

la différence normalisée des séries génératrices de la profondeur pour un arbre des suffixes et pour un trie définie pour $|u| < 1$. La série génératrice bivariée $Q(z,u)$ définie pour $|u| < 1$ et $|z| < 1$ est aussi introduite :

$$Q(z,u) := \sum_{n\geq 0} nQ_n(u)z^n = \sum_{w\in\mathcal{A}^\star} u^{|w|}p_w\left(\frac{z^{|w|}}{\mathfrak{D}_w^2(z)} - \frac{z}{(1-z(1-p_w))^2}\right) = \sum_{w\in\mathcal{A}^\star}\Psi_w(z,u).$$

Le théorème de Cauchy est appliqué à chaque $\Psi_w(z,u)$ avec z situé sur le cercle centré de rayon ϕ (cf. section précédente). Pour chaque motif de taille supérieure à K' il y a exactement deux singularités pour chaque Ψ_w à l'intérieur du contour : en $z = A_w$ et en $(1-p_w)^{-1}$. La seconde singularité est toujours à l'intérieur du cercle puisque la condition (5.24) implique que

$$p_w\phi < p_w\phi^k < (p\phi)^k\left(1 + (1+\phi)\frac{br}{d}\frac{1}{1-r\phi}\right) < \alpha(\phi-1) < (\phi-1), \tag{5.27}$$

pour un motif w de taille $k > K'$ et en prenant $\alpha < 1$. Donc $(1-p_w)^{-1}$ est plus petit que le rayon ϕ et les deux singularités sont dans le disque.

Il vient

$$\sum_{w\in\mathcal{A}^\star} I_{w,n}(\phi,u) := \sum_{w\in\mathcal{A}^\star} \frac{1}{2i\pi}\int_{|z|=\phi} \frac{\Psi_w(z,u)}{z^{n+1}}dz = \sum_{w\in\mathcal{A}^\star} \operatorname{Res}\left(\frac{\Psi_w(z,u)}{z^{n+1}};0\right)$$
$$+\sum_{k\geq K'}\sum_{w\in\mathcal{A}^k}\left(\operatorname{Res}\left(\frac{\Psi_w(z,u)}{z^{n+1}};A_w\right)+\operatorname{Res}\left(\frac{\Psi_w(z,u)}{z^{n+1}};\frac{1}{1-p_w}\right)\right)+\sum_{k<K'}\sum_{w\in\mathcal{A}^k}\Omega_{w,n}(\phi,u)$$
$$= nQ_n(u)+\sum_{k\geq K'}\sum_{w\in\mathcal{A}^k} u^{|w|}p_w\left(\operatorname{Res}\left(\frac{z^{|w|}}{z^{n+1}\mathfrak{D}_w^2(z)};A_w\right)+\operatorname{Res}\left(\frac{z}{z^{n+1}(1-z(1-p_w))^2};\frac{1}{1-p_w}\right)\right)$$
$$+\sum_{k<K'}\sum_{w\in\mathcal{A}^k}\Omega_{w,n}(\phi,u), \tag{5.28}$$

où $\Omega_{w,n}(\phi,u)$ est la somme des résidus de $\Psi_w(z,u)/z^{n+1}$ dans le disque épointé (*i.e.* dont on enlève l'origine) de rayon ϕ. La localisation des pôles est connue pour des motifs suffisamment longs mais pour les motifs plus courts nous n'avons pas d'information. La somme des $\Omega_{w,n}(\phi,u)$ sur les motifs de tailles inférieures à K' est de l'ordre de $O(\phi^{-n})$. La fonction $z/(1-z(1-p_w))^2$ est analytique en $z=A_w$ et donc ne contribue pas au résidu de $\Psi_w(z,u)/z^{n+1}$, cette partie peut donc être ignorée dans le calcul. De même, $1/(1-p_w)$ n'est pas un pôle de la fonction $1/\mathfrak{D}_w^2(z)$.

Les résidus en les A_w sont obtenus par un développement limité classique et l'utilisation des valeurs de B_w et E_w.

$$\operatorname{Res}\left(\frac{z^{k-(n+1)}}{\mathfrak{D}_w^2(z)};A_w\right) = A_w^{k-n-1}\left(\frac{k-(n+1)}{B_w^2A_w}-\frac{E_w}{B_w^3}\right). \tag{5.29}$$

Les résidus en les $z=(1-p_w)^{-1}$ se calculent en utilisant la formule générale $[z^n](1-az)^{-2}=(n+1)a^n$ qui amène

$$\left[z^{-1}\right]\frac{1}{z^n(1-z(1-p_w))^2} = \left[z^{n-1}\right]\frac{1}{(1-z(1-p_w))^2} = n(1-p_w)^{n-1}. \tag{5.30}$$

Avant de nous occuper du comportement asymptotique de la différence $Q_n(u)$, un lemme technique (identique au lemme 33 déjà utilisé et démontré page 108 du chapitre précédent) est nécessaire :

Lemme *Pour toute fonction f définie sur les motifs de $\mathcal{A}^\star$ et tout y réel,*

$$\sum_{w\in\mathcal{A}^k} p_w f(w) \leq \chi_k \mathbb{P}(\{w\in\mathcal{A}^k : f(w)>y\}) + y,$$

où χ_k est le maximum de la fonction f sur l'ensemble des motifs de taille k.

Il nous faut contrôler la croissance de l'intégrale $I_{w,n}(\phi,u)$ et de la somme des intégrales sur l'ensemble des motifs de taille k pour montrer que la somme converge. Pour chaque motif w de taille k, on écrit

$$\left|\frac{z^k}{\mathfrak{D}_w^2(z)}-\frac{z}{(1-z(1-p_w))^2}\right| = \left|\frac{z^k}{\mathfrak{D}_w^2(z)}-\frac{z^k}{(1-z(1-p_w))^2}+\frac{z^k}{(1-z(1-p_w))^2}-\frac{z}{(1-z(1-p_w))^2}\right|$$
$$\leq \left|\frac{z^k}{\mathfrak{D}_w^2(z)}-\frac{z^k}{(1-z(1-p_w))^2}\right|+\left|\frac{z^k}{(1-z(1-p_w))^2}-\frac{z}{(1-z(1-p_w))^2}\right|$$
$$\leq |z|^k\left|\frac{1}{\mathfrak{D}_w^2(z)}-\frac{1}{(1-z(1-p_w))^2}\right|+\frac{1}{|(1-z(1-p_w))^2|}\left|z^k-z\right|.$$

Cette majoration est valable pour n'importe quel z. Le premier terme de la majoration est noté $\Psi_{1,w}(z)$ et le second, $\Psi_{2,w}(z)$. On se place maintenant sur le cercle de rayon ϕ.

Aucun motif w n'a une probabilité arbitrairement proche de 1, donc il existe une constante $\mathfrak{K}_1$ pour laquelle tous les motifs de taille k vérifient

$$\left|\frac{1}{(1-z(1-p_w))^2}\right| \leq \mathfrak{K}_1.$$

La même constante $\mathfrak{K}_1$ est valable pour tous les motifs de taille supérieure à k, puisque la plus grande des probabilités p_w ne peut que diminuer quand la taille des motifs augmente ($p_w \leq p^{|w|}$). De plus $|z^k - z| \leq 2\phi^k$. Ainsi la majoration

$$\Psi_{2,w}(z) \leq 2\mathfrak{K}_1\phi^k$$

est valable pour tout entier k supérieur à une borne L_1.

La fonction $\Psi_{1,w}(z)$ se récrit

$$\left|\frac{(\mathfrak{D}_w(z)-(1-z(1-p_w))(\mathfrak{D}_w(z)+(1-z(1-p_w))}{(\mathfrak{D}_w(z)(1-z(1-p_w)))^2}\right|.$$

Le dénominateur ne peut devenir arbitrairement petit : les racines A_w des $\mathfrak{D}_w(z)$ sont de la forme $1 + p_w/c(1)$. Donc quand la taille des motifs devient grande, les pôles se rapprochent de 1 (car $A_w \leq 1 + p^k$) et par conséquent s'éloignent de ϕ. Il existe une constante $\mathfrak{K}_2$ pour laquelle

$$\left|\frac{1}{(\mathfrak{D}_w(z)(1-z(1-p_w)))^2}\right| \leq \mathfrak{K}_2$$

et ceci est vrai pour toutes les tailles k supérieures à L_2. Une majoration brutale permet d'écrire $\mathfrak{D}_w(z) + 1 - z(1-p_w) = O(\phi^k)$.

Les lemmes 41 et 33 servent à la majoration de

$$\mathfrak{D}_w(z) - (1-z(1-p_w)) = (1-z)(c_w(z)-1) + p_w(z^k(1+(1-z)F(z)) - z).$$

Notre objectif est d'obtenir une majoration de la somme

$$\sum_{w\in\mathcal{A}^k}\Psi_w(z,u) = \sum_{w\in\mathcal{A}^k} u^k p_w\left(\frac{z^k}{\mathfrak{D}_w^2(z)} - \frac{z}{(1-z(1-p_w))^2}\right),$$

sur $|z| = \phi$. Il est alors naturel de chercher une majoration sur le cercle de rayon ϕ de

$$\sum_{w\in\mathcal{A}^k} u^k p_w(\mathfrak{D}_w(z) - (1-z(1-p_w))).$$

Le lemme 41 affirme que

$$\sum_{w\in\mathcal{A}^k}[\![|c_w(\phi)-1| \leq (\phi\delta)^k\theta]\!]p_w \geq 1-\theta\delta^k,$$

pour un $\theta > 0$ et un $\delta < 1$ qui vérifie $\phi\delta < 1$ (quitte à prendre des bornes plus grandes sur la taille des motifs, on arrive à trouver un tel δ). Dès lors, après plusieurs majorations et en utilisant (5.23), il vient

$$\sum_{w\in\mathcal{A}^k}[\![\mathfrak{D}_w(z) - (1-z(1-p_w)) \leq (\phi\delta)^k\theta']\!]p_w \geq 1-\theta'\delta^k,$$

pour une constante $\theta' > 0$. Le lemme 33 s'applique avec $y = (\phi\delta)^k\theta'$ et $\chi_k = O(1)$.

Dans les paragraphes précédents, il fallait souvent que les motifs soient d'une taille supérieure à une certaine borne pour résoudre nos problèmes : K est le maximum de toutes ces conditions sur la longueur des motifs. Nous venons d'obtenir une borne sur la somme des intégrales $I_w(\phi, u)$ sur tous les motifs de taille k supérieure à K :

$$\sum_{w\in\mathcal{A}^k} I_{w,n}(\phi, u) = O((\delta\phi u)^k\phi^{-n}) \tag{5.31}$$

Les motifs de taille plus petite que K sont en nombre fini et ne contribuent dans leur ensemble qu'en $O(\phi^{-n})$. Pourvu que $|u|\delta\phi < 1$, la somme des $I_{w,n}(\phi, u)$ sur tous les motifs converge. Nous aboutissons ainsi au lemme suivant

Lemme 45 *Il existe $\beta > 1$ et $B > 1$ tels que pour tout $|u| \le \beta$ et pour n suffisamment grand, la différence $Q_n(u)$ s'écrit*

$$Q_n(u) = \frac{1}{n}\sum_{w\in\mathcal{A}^\star} u^{|w|}p_w\left(A_w^{|w|-n-1}\left(\frac{(n+1)-|w|}{B_w^2A_w} + \frac{E_w}{B_w^3}\right) - n(1-p_w)^{n-1}\right) + O(B^{-n}). \tag{5.32}$$

5.4.3 Comportement asymptotique de $Q_n(u)$

Lemme 46 *Pour tout β tel que $1 < \beta < \delta^{-1}$, il existe un c positif tel que la différence $Q_n(u) = O(n^{-c})$ uniformément pour tout $|u| \le \beta$.*

Preuve : Pour n suffisamment grand le terme dominant de l'equation (5.32) est

$$Q_n(u) = \sum_{w\in\mathcal{A}^\star} u^{|w|}p_w\left(\frac{A_w^{|w|-n-2}}{B_w^2} - (1-p_w)^{n-1}\right) + O(n^{-1}). \tag{5.33}$$

La fonction

$$f_w(z) = \left[\frac{A_w^{|w|-z-2}}{B_w^2} - (1-p_w)^{z-1}\right] - \left[\frac{1}{A_w^{2-|w|}B_w^2} - \frac{1}{1-p_w}\right]\exp(-z)$$

est au cœur de l'analyse de Mellin de la somme $f(z, u) := \sum_w u^{|w|}p_w f_w(z)$. Il aurait semblé plus naturel de s'occuper de la fonction de base

$$\frac{A_w^{|w|-z-2}}{B_w^2} - (1-p_w)^{z-1},$$

mais son comportement quand z tend vers zéro est assez délicat à obtenir. La fonction $f_w(z)$ quant à elle, se comporte en $O(z)$ au voisinage de l'origine.

La transformée de Mellin $f^\star(s)$ de $f(z)$ est bien définie dans la bande fondamentale $\langle -1, \infty\rangle$. L'application du lemme 33, avec $f(w) := u^{|w|}f_w(z)$, $y = (|u|\delta)^k$ pour un $\delta < 1$, $|u| \le \beta$, $\chi_k = O(1)$ et tout z montre que la somme $f(z, u)$ est absolument convergente (on distingue deux cas suivant si k est grand ou non). Donc

$$f_w^\star(s) = \Gamma(s)\left(\frac{(\log A_w)^{-s}}{A_w^{2-|w|}B_w^2} - \frac{(-\log(1-p_w))^{-s}}{1-p_w}\right) \text{ et } f^\star(s, u) = \sum_{w\in\mathcal{A}^\star} u^{|w|}p_w f_w^\star(s).$$

La transformée de Mellin $f^\star(s,u)$ est analytique dans la bande $-1 < \Re(s) < c$ pour un réel positif c. Pour prouver cela, on divise la somme $f^\star(s,u)$ en deux sous-sommes : la somme $\Xi_1(s,u)$ sur les motifs de $\mathcal{A}^\star$ qui vérifient $\Im(s)p_w$ petit et la somme $\Xi_2(s,u)$ sur les motifs pour lesquels $\Im(s)p_w$ est grand.

Pour chaque motif vérifiant $\Im(s)p_w$ petit, la fonction $f^\star_w(s)$ se développe en fonction de p_w pour s est fixé et on obtient

$$\begin{aligned}\frac{(\log A_w)^{-s}}{A_w^{2-|w|}B_w^2} - \frac{(-\log(1-p_w))^{-s}}{1-p_w} &= p_w^{-s}[(c_w(1))^{s-2}(1+O(p_w))(1+O(|w|p_w)) \\ &\quad -(1+O(p_w))(1+O(p_w))] \\ &= p_w^{-s}[(c_w(1))^{s-2}(1+O(|w|p_w)) - (1+O(p_w))].\end{aligned}$$

La valeur du polynôme d'auto-corrélation en 1 peut se récrire $c_w(1) = 1 + a_w(1)$, et ainsi

$$|c_w^{s-2}(1) - 1| \leq K(s)|c_w(1) - 1|,$$

avec $K(s)$, un constante fonction de s (s est fixe). Il suffit alors d'appliquer le lemme 33 avec $f(w) = |u|^k p^{-k\Re(s)} K(s)|c_w(1) - 1|$ pour obtenir

$$|\Xi_1(s)| \leq \sum_{k\geq 0} |u|^k p^{-k\Re(s)} \delta^k O(1). \tag{5.34}$$

La borne supérieure converge dès que $|u|p^{-\Re(s)}\delta < 1$. Or $|u| \leq \beta$, la condition sur s devient alors $\Re(s) < c$ pour c vérifiant $p^{-c}\beta\delta < 1$. Et puisque par hypothèse $\beta\delta < 1$, le réel c est strictement positif.

Pour un s fixé, il n'existe qu'un nombre fini de motifs w pour lesquels $\Im(s)p_w$ est grand. La contribution individuelle de chacun de ces motifs à la somme $f^\star(s,u)$ peut être grande mais puisqu'ils ne sont qu'en nombre limité, ils ne peuvent créer de singularité en s.

La transformée de Mellin $f^\star(s,u)$ est analytique dans la bande $-1 < \Re(s) < c$ donc le théorème de Mellin « inverse » nous garantit que $f(z,u) = O(z^{-c})$. Et par conséquent la différence $Q_n(u)$ est d'ordre $O(n^{-c})$ uniformément pour $|u| \leq \beta$ ◀

Ce lemme est suffisant pour arriver à notre objectif. Par définition $D_n^t(1) = D_n(1) = 1$, ainsi

$$Q_n(u) = \frac{u}{1-u}(D_n(u) - D_n^t(u)) = u\left(\frac{D_n(u) - D_n(1)}{1-u} - \frac{D_n^t(u) - D_n^t(1)}{1-u}\right). \tag{5.35}$$

De plus $D_n'(1) = \mathbb{E}_n(D_n)$ et $(D_n^t)'(1) = \mathbb{E}_n(D_n^t)$, ainsi quand u tend vers 1 dans $Q_n(u)$ (et $Q_n(u)$ est bien définie autour de 1) nous obtenons

$$\mathbb{E}_n(D_n^t) - \mathbb{E}_n(D_n) = O(n^{-c}). \tag{5.36}$$

Cela signifie qu'asymptotiquement la différence entre les deux moyennes (pour un trie et un arbre des suffixes) est de l'ordre de $O(n^{-c})$ pour un c positif. Or le théorème énoncé dans l'introduction de ce chapitre (à la page 133) donne le comportement asymptotique de la moyenne de la profondeur typique dans un trie sous modèle de Markov d'ordre 1. Nous pouvons ainsi conclure.

5.5 Conclusion

Nous venons de montrer que les moyennes de la profondeur typique dans un trie et dans un arbre des suffixes se comportent asymptotiquement de manière similaire pour une source de Markov d'ordre 1 jusqu'au terme d'ordre n^{-c} pour $c > 0$.

Théorème *Pour un arbre des suffixes construit sur les n premiers suffixes d'un texte produit par une source de Markov d'ordre 1, la moyenne de la profondeur typique se comporte asymptotiquement en*

$$\mathbb{E}_n(D_n) = \frac{1}{h}\left(\log n + \gamma + \frac{h_2}{2h} - H + \epsilon(n)\right) + O(n^{-c}),$$

où c est un réel positif, γ la constante d'Euler, h l'entropie de la source de Markov, h_2 la deuxième entropie, H l'entropie stationnaire et $\epsilon(z)$ une fonction de très faible amplitude fluctuant autour de zéro. En particulier

$$h := -\sum_{(i,j)\in\mathcal{A}^2} \pi_i p_{i,j} \log p_{i,j}, \quad H := -\sum_{i\in\mathcal{A}} \pi_i \log \pi_i \ \text{ et}$$

$$h_2 := h_2(1) \ \text{ avec } h_2(s) := \lim_{k\to\infty} \frac{d^2}{ds^2}\left(\sum_{w\in\mathcal{A}^k} p_w^s\right)^{1/k}.$$

Ce résultat s'étend à un modèle de Markov d'ordre fini quelconque. La seule difficulté réside dans le grand nombre d'indices dans les expressions.

Une analyse étendue devrait permettre d'obtenir les résultats sur le comportement asymptotique de la variance et la distribution limite de la profondeur d'un arbre des suffixes sous un modèle de source markovienne d'ordre 1. Le paramètre devrait, comme dans le cas d'une source sans mémoire suivre une distribution normale. La plus grande précision des outils utilisés dans ce chapitre permet d'améliorer les résultats du chapitre 3 sur la taille et la longueur de cheminement.

Comme pour les paramètres des chapitres précédents, l'étude de la profondeur typique devra être réalisée pour la classe plus large des sources dynamiques introduites par Vallée dans [Val01].

Ce chapitre se focalise sur la longueur des phrases dans le découpage d'un texte par l'algorithme LZ'77. La longueur de la plus longue phrase est en fait la hauteur de l'arbre des suffixes construit sur le texte. Plusieurs autres paramètres liés au découpage en phrases dans l'algorithme LZ'77 sont intéressants : le nombre de textes de taille n avec m phrases, le nombre de phrases complètes dans un texte de taille n et la taille d'un texte composé de m phrases. Ces paramètres ont déjà fait l'objet de plusieurs études pour l'algorithme-frère LZ'78 [GK92, JS95, LS95, JST01]. Les méthodes utilisées dans ce chapitre devraient permettre d'obtenir certaines estimations de ces quantités dans les arbres des suffixes.

Perspectives

Plusieurs recherches futures ont été mentionnées dans les paragraphes de conclusion des chapitres 2, 3, 4 et 5.

Ces remarques concernent d'abord l'extension des résultats à des modèles plus puissants de génération des symboles. La moyenne de la taille et de la longueur de cheminement dans un arbre des suffixes ont été obtenues dans le chapitre 2 pour des textes engendrés par une source sans mémoire avec une forte contrainte sur les valeurs des probabilités. Le résultat du chapitre 5 sur la moyenne de la profondeur typique, montre qu'il est possible d'étendre les résultats du chapitre 2 au moins à un modèle markovien.

L'objectif initial de ma thèse était d'obtenir des résultats sur le comportement asymptotique de la moyenne de la taille et de la longueur de cheminement dans un arbre des suffixes pour un texte engendré par une source dynamique (cf. [Val01]). Les jalons de la méthode ont été posés dans le chapitre 2 pour la taille et la longueur de cheminement : obtenir une expression de la moyenne des ces paramètres pour un arbre des suffixes et comparer le comportement asymptotique de la moyenne de ces deux paramètres dans un trie et dans un arbre des suffixes. Dans [CFV01], Clément, Flajolet et Vallée ont obtenu les résultats pour la moyenne de la taille et de la longueur de cheminement d'un trie sous un modèle de source dynamique. Certaines études préliminaires pour les arbres des suffixes sous un modèle de source dynamique ont été menées dans le cadre de ma thèse mais ne sont pas incluses dans ce manuscript. Des obstacles demeurent pour obtenir le résultat initialement souhaité mais semblent raisonnablement pouvoir être levés.

Un autre champ d'étude est le passage des résultats en moyenne obtenus (ou en variance comme au chapitre 3) à des résultats sur la distribution des paramètres étudiés.

Pour la variance du chapitre 3, le point d'achoppement est pour l'instant la détermination des pôles dominants des séries génératrices de la section 3.4. Ensuite nous devrons comparer les variances de la taille et de la longueur de cheminement d'un trie construit sur n chaînes avec celles d'un arbre des suffixes construit sur les n premiers suffixes d'un texte.

Le polynôme d'auto-corrélation joue un rôle central dans cet thèse. En collaboration avec Jérémie Bourdon, nous avons entamé un travail, dans le prolongement de ceux de Guibas et Odlyzko [GO81a], visant à améliorer le comportement asymptotique du nombre de polynomes d'auto-corrélation (identifiés par leurs coefficients 0 ou 1) pour des textes de taille n donnée. Ce travail n'est pas inclus dans la thèse.

Il existe des liens entre d'une part le nombre de nœuds internes et le niveau de saturation dans un trie et dans un arbre des suffixes et d'autre part des versions étendues du paradoxe des anniversaires et du problème du collecteur de coupons. J'ai commencé à étudier plusieurs questions intéressantes qui y sont reliées : combien de lettres faut-il en moyenne pour qu'un texte contienne tous les mots de taille k si les mots sont comptés sans se chevaucher ? et si les mots sont comptés en se chevauchant ? Combien de lettres faut-il en moyenne dans le texte pour avoir exactement j motifs de taille k qui apparaissent au moins à m reprises ? On peut remarquer

que si $m = 2$, la question revient à : combien de suffixes successifs faut-il considérer pour avoir exactement j nœuds internes à profondeur k ?

D'autres travaux sont en cours sur la localisation de sites promoteurs sur le génome humain ainsi que sur le comptage de co-occurrences de certains mots.

Bibliographie

[AN93] Arne Andersson and Stefan Nilsson. Improved behaviour of tries by adaptive branching. *Information Processing Letters*, 46 :295–300, 1993.

[Apo85] Alberto Apostolico. The myriad virtues of subword trees. In A. Apostolico and Z. Galil, editors, *Combinatorial Algorithms on Words*, volume 12 of *NATO Advance Science Institute Series.* Series F : Computer and Systems Sciences, pages 85–96. Springer Verlag, 1985.

[AS92] Alberto Apostolico and Wojciech Szpankowski. Self-alignements in words and their applications. *Journal of Algorithms*, 13 :446–467, 1992.

[BCRV05] Valentina Boeva, Julien Clément, Mireille Régnier, and Mathias Vandenbogaert. Assessing the significance of sets of words. In *Combinatorial Pattern Matching 05*, volume 3537 of *Lecture Notes in Computer Science*, pages 358–370. Springer Verlag, 2005.

[BEH89] Anselm Blumer, Andrezj Ehrenfeucht, and David Haussler. Average sizes of suffix trees and dawgs. *Discrete Applied Mathematics*, 24(1) :37–45, 1989.

[Bel86] Timothy C. Bell. Better OPM/L text compression. *IEEE Transactions on Computers*, 34 :1176–1182, 1986.

[BHJ92] Andrew D. Barbour, Lars Holst, and Svante Janson. *Poisson approximation.* The Clarendon Press Oxford University Press, New York, 1992. Oxford Science Publications.

[BYN00] Ricardo Baeza-Yates and Gonzalo Navarro. A hybrid indexing method for approximate string matching. *Journal of Discrete Algorithms*, 1(1) :205–239, 2000.

[CFV01] Julien Clément, Philippe Flajolet, and Brigitte Vallée. Dynamical sources in information theory : A general analysis of trie structures. *Algorithmica*, 29(1/2) :307–369, 2001.

[Clé00] Julien Clément. *Arbres digitaux et sources dynamiques.* Thèse de doctorat, Université de Caen, September 2000.

[CMRS99] Maxime Crochemore, Filippo Mignosi, Antonio Restivo, and Sergio Salemi. Text compression using antidictonaries. In J. Wiedermann, P. van Emde Boas, and M. Nielsen, editors, *International Conference on Automata, Languages an Programming (Prague, 1999)*, volume 1644 of *LNCS*, pages 261–270. Springer-Verlag, 1999.

[CMRS00] Maxime Crochemore, Filippo Mignosi, Antonio Restivo, and Sergio Salemi. Data compression using antidictonaries. In J. Storer, editor, *Proceedings of the I.E.E.E., Lossless Data Compression*, pages 1756–1768, 2000.

[CN02] Maxime Crochemore and Gonzalo Navarro. Improved antidictionary based compression. In *SCCC'02, Chilean Computer Science Society*, pages 7–13. I.E.E.E. CS Press, Nov. 2002.

[dBKR72] N. G. de Bruijn, D. E. Knuth, and S. O. Rice. The average height of planted plane trees. In R. C. Read, editor, *Graph Theory and Computing*, pages 15–22. Academic Press, 1972.

[Dev82] Luc Devroye. A note on the average depth of tries. *Computing*, 28 :367–371, 1982.

[Dev92] Luc Devroye. A study of trie-like structures under the density model. *Annals of Applied Probability*, 2 :402–434, 1992.

[Dev01] Luc Devroye. Analysis of random LC-tries. *Random Structures & Algorithms*, 15 :359–375, 2001.

[Dev02] Luc Devroye. Laws of large numbers and tail inequalities for random tries and Patricia trees. *Journal of Computational and Applied Mathematics*, 142 :27–37, 2002.

[dlB59] René de la Briandais. File searching using variable length keys. In *Proceedings, Western Joint Computer Conference*, pages 295–298, 1959.

[DSR92] Luc Devroye, Wojciech Szpankowski, and Bonita Rais. A note on the height of suffix trees. *SIAM Journal on Computing*, 21(1) :48–53, 1992.

[Fay04] Julien Fayolle. An average-case analysis of basic parameters of the suffix tree. In Michael Drmota, Philippe Flajolet, Danièle Gardy, and Bernhard Gittenberger, editors, *Mathematics and Computer Science*, pages 217–227. Birkhäuser, 2004. Proceedings of a colloquium organized by TU Wien, Vienna, Austria, September 2004.

[Fel68] William Feller. *An Introduction to Probability and Its Application*. John Wiley and Sons, New York, third edition, 1968.

[FFH86] Guy Fayolle, Philippe Flajolet, and Micha Hofri. On a functional equation arising in the analysis of a protocol for a multiaccess broadcast channel. *Advances in Applied Probability*, 18 :441–472, 1986.

[FFMO06] Julien Fayolle, Philippe Flajolet, Hiroyoshi Morita, and Takahiro Ota. Average size of the antidictionary in DCA. to appear, 2006.

[FGD95] Philippe Flajolet, Xavier Gourdon, and Philippe Dumas. Mellin transforms and asymptotics : Harmonic sums. *Theoretical Computer Science*, 144(1–2) :3–58, June 1995.

[FGT92] Philippe Flajolet, Danièle Gardy, and Loÿs Thimonier. Birthday paradox, coupon collectors, caching algorithms, and self–organizing search. *Discrete Applied Mathematics*, 39 :207–229, 1992.

[Fla83] Philippe Flajolet. On the performance evaluation of extendible hashing and trie searching. *Acta Informatica*, 20 :345–369, 1983.

[FPS96] Ioannis Fudos, Evaggelia Pitoura, and Wojciech Szpankowski. On pattern occurrences in a random text. *Information Processing Letters*, 57 :307–312, 1996.

[Fre60] Edward H. Fredkin. Trie memory. *Communications of the ACM*, 3 :490–499, 1960.

[FS86] Philippe Flajolet and Robert Sedgewick. Digital search trees revisited. *SIAM Journal on Computing*, 15(3) :748–767, August 1986.

[FS06] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. 2006.

[FW05] Julien Fayolle and Mark Daniel Ward. Analysis of the average depth in a suffix tree under a markov model. In *Proceedings of the 2005 International Conference on the Analysis of Algorithms*, pages 95–104. DMTCS, 2005. Proceedings of a colloquium organized by Universitat Politècnica de Catalunya, Barcelona, Catalunya, June 2005.

[GJ83] Ian P. Goulden and David M. Jackson. *Combinatorial enumeration*. John Wiley & Sons, 1983.

[GK92] Edgar Gilbert and T. T. Kadota. The Lempel-Ziv algorithm and message complexity. *IEEE Transaction on Information Theory*, 38 :1839–1842, 1992.

[GK97] Robert Giegerich and Stefan Kurtz. From ukkonen to mccreight and weiner : A unifying view of linear-time suffix tree construction. *Algorithmica*, 19 :331–353, 1997.

[GO81a] Leo J. Guibas and Andrew M. Odlyzko. Periods in strings. *Journal of Combinatorial Theory,* Series A, 30 :19–42, 1981.

[GO81b] Leo J. Guibas and Andrew M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A*, 30(2) :183–208, 1981.

[Gus97] Daniel Gusfield. *Algorithm on Strings, Trees, and Sequences : Computer Science and Computational Biology.* Cambridge Press, 1997.

[HAI01] Ela Hunt, Malcolm P. Atkinson, and Robert W. Irving. A database index to large biological sequences. In *Proceedings of the 27th Conference on Very Large Databases, Rome*, pages 139–148, 2001.

[JR86] Philippe Jacquet and Mireille Régnier. Trie partitioning process : Limiting distribution. In Springer-Verlag, editor, *Lecture Notes in Computer Science*, volume 214, pages 196–210, 1986.

[JR87] Philippe Jacquet and Mireille Régnier. Normal limiting distribution of the size of tries. In North-Holland, editor, *Proc. Performance '87*, pages 209–223, 1987.

[JR88] Philippe Jacquet and Mireille Régnier. Normal limiting distribution of the size and the external path length of tries. Technical report, Institut National de Recherche en Informatique et en Automatique, 1988.

[JS91] Philippe Jacquet and Wojciech Szpankowski. Analysis of digital tries with markovian dependency. *IEEE Transactions on Information Theory*, 37(5) :1470–1475, 1991.

[JS94] Philippe Jacquet and Wojciech Szpankowski. Autocorrelation on words and its applications : analysis of suffix trees by string-ruler approach. *Journal of Combinatorial Theory. Series A*, 66(2) :237–269, 1994.

[JS95] Philippe Jacquet and Wojciech Szpankowski. Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees. *Theoretical Computer Science*, 144(1–2) :161–197, 1995.

[JS05] Philippe Jacquet and Wojciech Szpankowski. *Applied Combinatorics on Words*, chapter Analytic Approach to Pattern Matching. Cambridge University Press, 2005.

[JST01] Philippe Jacquet, Wojciech Szpankowski, and Jing Tang. Average profile of the Lempel-Ziv parsing scheme for a markovian source. *Algorithmica*, 31 :318–360, 2001.

[Kac49] Marek Kac. On the deviations between theoretical and empirical distributions. *Proceedings of the National Academy of Science USA*, 35 :252–257, 1949.

[Knu73] Donald E. Knuth. *The Art of Computer Programming*, volume 3 : Sorting and Searching. Addison-Wesley, 1973.

[KP88] Peter Kirschenhofer and Helmut Prodinger. Further results on digital search trees. *Theoretical Computer Science*, 58 :143–154, 1988.

[KP91] Peter Kirschenhofer and Helmut Prodinger. On some applications of formulæ of Ramanujan in the analysis of algorithms. *Mathematika*, 38 :14–33, 1991.

[KPS89] Peter Kirschenhofer, Helmut Prodinger, and Wojciech Szpankowski. On the variance of the external path length in a symmetric digital trie. *Discrete Applied Mathematics*, 25 :129–143, 1989.

[KS60] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. van Nostrand, 1960.

[LS95] Guy Louchard and Wojciech Szpankowski. Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm. *IEEE Transactions on Information Theory*, 41 :478–488, 1995.

[LS03] Stefano Lonardi and Wojciech Szpankowski. Joint source-channel lz'77 coding. In *2003 Data Compression Conference (DCC 2003)*, pages 273–282, 2003.

[LST99] Guy Louchard, Wojciech Szpankowski, and Jing Tang. Average profile of the generalized digital search tree and the generalized lempel-ziv algorithm. *SIAM J. Computing*, 28 :935–954, 1999.

[LSW04] Stefano Lonardi, Wojciech Szpankowski, and Mark Daniel Ward. Error resilient lz'77 scheme and its analysis. In *2004 IEEE International Symposium on Information Theory*, 2004.

[Mah92] Hosam M. Mahmoud. *Evolution of Random Search Trees*. John Wiley, New York, 1992.

[Mar83] Hugo M. Martinez. An efficient method for finding repeats in molecular sequences. *Nucleic Acids Research*, 11 :4629–4634, 1983.

[McC76] Edward McCreight. A space-economical suffix tree construction algorithm. *Journal of the ACM*, 23(2) :262–272, 1976.

[MO04] Hiroyoshi Morita and Takahiro Ota. An upper bound on size of antidictionary. In *Proceedings of SITA2004*, 2004.

[Mor68] Donald Morrison. Patricia–practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM*, 15(4) :514–534, 1968.

[Nic03] Pierre Nicodème. Q-gram analysis and urn models. In Cyril Banderier, editor, *Proceedings of Discrete Random Walks*, pages 25–30, 2003.

[NSF99] Pierre Nicodème, Bruno Salvy, and Philippe Flajolet. Motif statistics. In J. Nešetril, editor, *Algorithms, ESA'99*, volume 1643 of *Lecture Notes in Computer Science*, pages 194–211, 1999.

[OM04] Takahiro Ota and Hiroyoshi Morita. One-path ECG lossless compression using antidictionaries. *IEICE Trans. Fundamentals (Japanese Edition)*, J87-A(9) :1187–1195, 2004.

[Pit85] Boris Pittel. Asymptotical growth of a class of random trees. *Annals of Probability*, 13(2) :414–427, 1985.

[Pit86] Boris Pittel. Paths in a random digital tree : Limiting distributions. *Advances in Applied Probability*, 18 :139–155, 1986.

[RD04] Mireille Régnier and Alain Denise. Rare events and conditional events on random strings. *Discrete Mathematics and Theoretical Computer Science*, 6 :191–214, 2004.

[RJ89] Mireille Régnier and Philippe Jacquet. New results on the size of tries. *IEEE Transactions on Information Theory*, 35(1) :203–205, 1989.

[RS97] Mireille Régnier and Wojciech Szpankowski. On the approximate pattern occurrences in a text. In IEEE Computer Society, editor, *Compression and Complexity of SEQUENCES 1997*, pages 253–264, 1997. In Proceedings SEQUENCE'97, Positano, Italy.

[RS98] Mireille Régnier and Wojciech Szpankowski. On pattern frequency occurrences in a markovian sequence. *Algorithmica*, 22(4) :631–649, 1998. This paper was presented in part at the 1997 *International Symposium on Information Theory*, Ulm, Germany.

[Sch00] Sophie Schbath. An overview on the distribution of word counts in Markov chains. *Journal of Computational Biology*, 7 :193–201, 2000.

[SS82] James A. Storer and Thomas G. Szymanski. Data compression via textual substitution. *Journal of the ACM*, 29 :928–951, 1982.

[Szp88] Wojciech Szpankowski. Some results on V-ary asymmetric tries. *Journal of Algorithms*, 9 :224–244, 1988.

[Szp91] Wojciech Szpankowski. On the height of digital trees and related problems. *Algorithmica*, 6(2) :256–277, 1991.

[Szp93a] Wojciech Szpankowski. Asymptotic properties of data compression and suffix trees. *IEEE Transactions on Information Theory*, 39 :1647–1659, 1993.

[Szp93b] Wojciech Szpankowski. A generalized suffix tree and its (un)expected asymptotic behaviors. *SIAM Journal on Computing*, 22 :1176–1198, 1993.

[Szp01] Wojciech Szpankowski. *Average-Case Analysis of Algorithms on Sequences*. John Wiley, New York, 2001.

[THP04] Sandeep Tata, Richard A. Hankins, and Jignesh M. Patel. Practical suffix tree construction. In *Proceedings of the 30th Very Large Data Base Conference (Toronto)*, 2004.

[Ukk95] Esko Ukkonen. On-line construction of suffix-trees. *Algorithmica*, 14 :249–260, 1995.

[Val01] Brigitte Vallée. Dynamical sources in information theory : Fundamental intervals and word prefixes. *Algorithmica*, 29(1/2) :262–306, 2001.

[Wat95] Michael S. Waterman. *Introduction to Computational Biology*. Chapman & Hall/CRC, 1995.

[Wei73] Peter Weiner. Linear pattern matching algorithm. In *IEEE 14th Annual Symposium on Switching and Automata Theory*, pages 1–11, 1973.

[WS04] Mark Daniel Ward and Wojciech Szpankowski. Analysis of a randomized selection algorithm motivated by the LZ'77 scheme. In *The First Workshop on Analytic Algorithmics and Combinatorics (ANALCO 04)*, 2004.

[WS05] Mark Daniel Ward and Wojciech Szpankowski. Analysis of the multiplicity matching parameter in suffix trees. In Discrete Mathematics and Theoretical Computer Science, editors, *2005 International Conference on the Analysis of Algorithms*, pages 307–322, 2005.

[Wyn97] Abraham J. Wyner. The redundancy and distribution of the phrase lengths of the fixed-database Lempel-Ziv algorithm. *IEEE Transactions on Information Theory*, 43(5) :1452–1464, September 1997.

[WZ89] Abraham J. Wyner and Jacob Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35 :1250–1258, 1989.

[ZL77] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23 :337–343, 1977.