UNIVERSITÉ MOULOUD MAMMERI DE TIZI-OUZOU
FACULTÉ DES SCIENCES
DÉPARTEMENT DE MATHÉMATIQUES

THÈSE DE DOCTORAT

Discipline : Mathématiques
Option : Statistique

Présentée
par

FERHAT ZIRAM

Soutenue publiquement le 07 avril 2016

Titre :

ANALYSE ET MODÉLISATION STATISTIQUE DES RÉPÉTITIONS CHEVAUCHANTES MAXIMALES À GAUCHE DANS DES SÉQUENCES SUR UN ALPHABET FINI. MÉTHODE DE CHEN-STEIN ET APPROXIMATION DE POISSON.

Devant le jury composé par :

M^r Morsli Mohamed	Professeur	UMMTO	Président
M^r Charlot François	Professeur	Université de Rouen (France)	Rapporteur
M^r Cellier Dominique	Professeur	Université de Rouen (France)	Examinateur
M^r Aïssani Djamil	Professeur	Université de Béjaïa	Examinateur
M^r Fellag Hocine	Professeur	UMMTO	Examinateur

À la mémoire de ma mère,
À mon père,
À ma femme et mes enfants,
À mes frères et sœurs.

Remerciements

Je tiens à remercier Monsieur François CHARLOT Professeur à l'université de Rouen (France) d'avoir accepté de diriger le travail de cette thèse.

Je tiens à remercier Monsieur Mohamed MORSLI Professeur à l'U.M.M.T.O d'avoir accepté d'être président de ce jury.

Je tiens à exprimer ma profonde gratitude à Monsieur Dominique CELLIER Professeur à l'université de Rouen (France) pour son aide précieuse le long de ce travail, et pour les facilités que j'ai trouvées auprès de lui au LITIS, et je le remercie d'avoir accepté de faire parti de ce jury.

Je tiens à remercier aussi Monsieur Djamil AÏSSANI Professeur à l'université de Béjaïa et Monsieur Hocine FELLAG Professeur à l'U.M.M.T.O d'avoir bien voulu faire parti de ce jury.

Un grand merci à toutes et à tous ceux qui m'ont encouragé de près ou de loin le long de ce travail.

Table des matières

Introduction

Dans de nombreux domaines des sciences expérimentales, l'étude d'un phénomène exige d'effectuer des observations sur de longues séries. Leur traitement et leur analyse nécessitent d'une part l'utilisation de l'outil informatique, qui grâce à des algorithmes performants, permet d'enregistrer les séquences et d'effectuer les calculs sur ces séquences, d'autre part l'outil statistique pour l'étude de la signaficativité statistique afin d'expliquer et interpréter le phénomène étudié. Nous citons à titre d'exemple dans le domaine de la biologie, le décodage du génome humain. Une étude expérimentale fiable nécessite de considérer des séquences très longues (c'est à dire de longues séries formées de quatre nucléotides `a`, `g`, `c` et `t` qui constituent l'ADN), ce qui est difficile, voire impossible à réaliser de manière manuelle. Parmi les méthodes utilisées, il y a celles qui consistent à comparer entres elles des séquences d'ADN de même longueur, afin de distinguer les ressemblances et les dissemblances, celles qui consistent à reconstituer une séquence d'ADN à partir de morceaux qui figurent dans la séquence, cette méthode est appelée séquençage. Dans les deux cas on a recours à la *détection* et à la *comparaison* de *répétitions significatives*, c'est à dire à s'intéresser à la nature, au nombre et la longueur, des répétitions. Pour cela, il est nécessaire de modéliser les séquences à comparer aux séquences observées par des séquences aléatoires.

L'idée de l'étude statistique des *répétitions chevauchantes maximales à gauche* (les mots répétés ne peuvent être étendus à d'autres de longueur plus grande) de mots de longueur donnée, est motivée par le travail de R.Arratia *and al.* (1996) [3] sur le séquençage par hybridation (sequencing by hybridisation SBH), où il s'agit de relier entre eux des fragments formés des 4 nucléotides `a`, `g`, `c` et `t` non ordonnés appelés *mots* de longueur t, afin de reconstituer la séquence d'ADN "exacte" qui est la séquence observée, de longueur n. La séquence à reconstituer est donc extraite d'une suite de variables aléatoires, autrement dit, elle est générée par une suite de variables aléatoires à valeurs sur un *alphabet fini* qui est dans ce cas l'ensemble $\{\mathtt{a}, \mathtt{g}, \mathtt{c}, \mathtt{t}\}$. Lors de cette reconstitution il peut y avoir des *répétitions chevauchantes maximales à gauche* et des *répétitions non chavauchantes maximales à gauche* de mots de même longueur t. Dans leur travail, R.Arratia *and al.* (1996) [3] se sont intéressés à l'approximation de la loi de probabilité du nombre N_t de répétitions (chevauchantes et non chevauchantes) maximales à gauche dans la séquence, où les variables aléatoires considérées sont indépendantes identiquement distribuées (i.i.d). Ils ont montré, en utilisant la méthode de Chen-Stein, que sous certaines conditions, la loi de probabilité de N_t est approximée par une loi de Poisson de paramètre λ, où $\lambda = \mathbb{E}(N_t)$ est le nombre moyen de répétitions maximales à gauche.

La détection des répétitions est donc d'un grand intérêt pour comparer une séquence observée à une séquence générée aléatoirement. Le cas (i-i-d) comme nous venons de le citer dans le paragraphe ci-dessus a été traité par R.Arratia *and al.* (1996) [3]. Dans le cas où la séquence aléatoire est extraite d'une chaîne de Markov homogène et stationnaire sur un alphabet fini (qui est un cas où les variables aléatoires ne sont pas indépendantes), N.Touyyar *and al,* (2008) [15] ont montré, en utilisant la méthode de Chen-Stein, que sous certaines conditions, la loi de probabilité du nombre de répétitions non chevauchantes maximales à gauche N_t est approximée par la loi Poisson de paramètre λ, où $\lambda = \mathbb{E}(N_t)$.

Dans notre travail, nous nous intéressons à l'analyse statistique des *répétitions chevauchantes maximales à gauche*, qui est dans notre cas l'étude de la significativité statistique des répétitions chevauchantes maximales à gauche. Ceci nécessite de calculer la p-value $\mathbb{P}(N_t \geq N_t^{obs})$, c'est à dire la probabilité qu'il y ait autant de répétitions chevauchantes maximales à gauche (dont le nombre est N_t^{obs}) dans la séquence observée que dans la séquence aléatoire, quand celle-ci est extraite d'une chaîne de Markov homogène et stationnaire sur un alphabet fini (dont le nombre est N_t). Tout comme dans le cas N.Touyyar *and al,* (2008) [15], nous montrons que sous certaines conditions, la méthode de Chen-Stein donne aussi l'approximation de la loi de probabilité de N_t par une loi de Poisson de paramètre λ, où $\lambda = \mathbb{E}(N_t)$. Ce qui permet donc de calculer ou du moins de donner une approximation de la valeur de la p-value.

Les premiers travaux se rapportant aux répétitions chevauchantes maximales à gauche, ont été réalisés pour l'étude de la fréquence des mots rares par G.Reinert et S.Schbath, (1998) [7], et les familles de mots rares par R.Roquin et S.Schbath, (2007) [9], lorsque la séquence aléatoire est modélisée par une chaîne de Markov homogène et stationnaire sur un alphabet fini. Dans les deux cas la loi de probabilité du nombre de répétitions est approximée par une loi de Poisson composée et non par une loi de Poisson. Contrairement aux cas traités par G.Reinert et S.Schbath, (1998) [7] et R.Roquin et S.Schbath, (2007) [9], nous considérons dans notre travail, les répétitions chevauchantes maximales à gauche sans tenir compte de la nature des mots répétés. Une répétition chevauchante maximale à gauche dans une séquence dans notre cas, est représentée par une position, qui est un couple de nombres entiers naturels non nuls et différents, dont la valeur absolue de la différence est strictement inférieure à la longueur du mot répété. Compter le nombre de répétitions dans la séquence revient donc à compter le nombre de positions de ces répétitions dans la séquence.

Il s'agit dans cette thèse d'analyse statistique des répétitions chevauchantes et maximales à gauche, lorsque les séquences sont modélisées par une chaîne de Markov homogène et stationnaire sur un alphabet fini. En plus de l'introduction que nous venons de développer ci-dessus sur les répétitions en général, et les répétitions chevauchantes maximales à gauche en particulier, où nous avons situé notre travail par rapport aux travaux de recherches réalisés, le manuscrit de cette thèse est composée essentiellement de trois chapitres et une conclusion.

Le premier chapitre comporte les définitions et les notations du modèle qui sont utilisées dans tout le travail. Nous définissons de manière générale une répétition d'un mot de longueur t en une position $\alpha = (i, j)$ (où i et j sont entiers naturels tels que $i \neq j$) dans une séquence S de longueur n, et en particulier celle d'une répétition chevauchante maximale à gauche d'un mot de longueur t, celle-ci étant de par sa définition un mot W_α de longueur $t + \ell$ où

$\ell = |i - j|$ avec $\ell = 1, \ldots, t-1$. Nous donnons la caractérisation d'une répétition chevauchante maximale à gauche d'un mot de longueur t, où nous montrons qu'une répétition chevauchante maximale à gaughe d'un mot de longueur t, est en fait la répétition des d (d étant un diviseur de ℓ) premières lettres du mot répété (d est alors appelé *période* de la répétition). Nous présentons ensuite le modèle markovien des séquences aléatoires dont les valeurs sont dans un alphabet fini $\mathcal{A}$, tout en donnant là aussi les définitions et les notations qui sont utilisées dans ce manuscrit. Nous déterminons pour ce modèle l'expression qui calcule le nombre N_t de répétitions chevauchantes maximales à gauche de mots de longueur t dans la séquence S, ainsi que celle du nombre moyen $\lambda = \mathbb{E}(N_t)$ de ces répétitions, qui dépend des probabilités de transition d'une lettre à une autre dans la répétition. En majorant chaque probabilité de transition de l'expression de λ par ξ $(0 < \xi < 1)$, nous obtenons l'ordre de grandeur de λ, qui pour $t = o(n)$ est majoré par $n^2\xi^t$, et nous montrons aussi que sous la condition $n^2\xi^t = O(1)$, le paramètre λ est borné sur $]0, +\infty[$.

Dans le deuxième chapitre, nous donnons la borne de Chen-Stein, qui est la borne supérieure de la distance en variation totale $d_{VT}(\mathcal{L}(N_t), \mathcal{P}_\lambda)$ entre la loi de probabilité $\mathcal{L}(N_t)$ de N_t et la loi de Poissson $\mathcal{P}_\lambda$ de paramètre λ, où $\lambda = \mathbb{E}(N_t)$. Il est montré dans R.Arratia *and al.* (1989) [1], que la borne de Chen-Stein fait intervenir trois quantités b_1, b_2 et b_3, plus exactement que la distance en variation totale $d_{VT}(\mathcal{L}(N_t), \mathcal{P}_\lambda)$ est majorée par $2(b_1 + b_2 + b_3)$, où b_1, b_2 et b_3 sont des sommes de toutes les occurrences de répétitions en positions α et β *voisines*. Ces répétitions peuvent être (particulièrement dans les expressions de b_1 et b_2) chevauchantes entre elles, autrement dit, il y a recouvrement entre les deux répétitions (la répétition en β recouvre la répétition en α si elle est précédée par celle-ci ou la répétition en α recouvre celle en β dans le cas contraire). Les répétitions dont il est question dans ce travail étant chevauchantes et maximales à gauche, il est donc nécessaire de vérifier la maximalité à gauche des répétitions quand il y a recouvrement entre elles. Ce qui nous conduit à donner dans la proposition 2.2.1, le corollaire 2.2.1 et le corllaire 2.2.2, des conditions suffisantes de maximalité à gauche pour de telles répétitions.

Le résultat principal de notre travail est le théorème 2.2.1, qui montre que la quantité $2(b_1 + b_2 + b_3)$, et donc $d_{VT}(\mathcal{L}(N_t), \mathcal{P}_\lambda)$, est majorée par une expression dépendant de $n^2\xi^t$, et que si $n^2\xi^t = O(1)$, la borne de Chen-Stein converge vers 0 quand la longueur n de la séquence S est assez grande, par conséquent la loi de probabilité $\mathcal{L}(N_t)$ est approximée par la loi de Poisson $\mathcal{P}_\lambda$.

La démonstration du théorème 2.2.1 repose sur la majoration des quantités b_1, b_2 et b_3. Les majoration b_1 et b_3 découlent directement de celles des sommes des probabilités des occurrences données dans leurs expressions, tandis que la majoration de b_2 dépend du recouvrement qu'il y a entre les répétitions en α et en β. Le recouvrement des ces répétitions forme un mot $\mathcal{W}_{\alpha,\beta}$ dont la longueur en dépend, ainsi majorer b_2 revient à majorer la somme des probabilités que le mot $\mathcal{W}_{\alpha,\beta}$ occure pour toute position β voisine de α avec $\alpha \neq \beta$. De ces recouvrement (nous supposons par symétrie que la répétition en α précède celle en β), nous déduisons deux principaux cas, celui où la longueur $|\mathcal{W}_{\alpha,\beta}|$ du mot $\mathcal{W}_{\alpha,\beta}$ est supérieure strictement à $2t$ (il y a 5 cas dont un ($k = 0$) correspondant à la longueur maximale de $\mathcal{W}_{\alpha,\beta}$, où les calculs sont déterminés de façon explicite, ce qui n'est pas le cas des autres), et celui où la longueur $|\mathcal{W}_{\alpha,\beta}|$ est inférieure strictement à $2t$ (il y a un seul cas parmi 3, les autres sont exclus du fait que d'après les conditions suffisantes de maximalité à gauche, la répétition en β n'est pas maximale à gauche). Notons aussi que le cas $|\mathcal{W}_{\alpha,\beta}| = 2t$ n'est pas pris en considération, du fait qu'il

est réalisé uniquement si la répétition en β n'est pas maximale à gauche. La quantité b_2 est donc décomposée suivant la longueur $|\mathcal{W}_{\alpha,\beta}|$ du mot $\mathcal{W}_{\alpha,\beta}$, en la somme de deux quantités $b_{2,\{|\mathcal{W}_{\alpha,\beta}|>2t\}}$ et $b_{2,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$, dont les majorations entrainent celle de b_2.

Dans le troisième chapitre, il est question d'élargir le résultat du chapitre 2 (le théorème 2.2.1) au cas où la séquence aléatoire S est générée par une chaîne de Markov d'ordre m (où m est un entier naturel supérieur ou égal à 1) noté modèle Mm. Ce type de modélisation est souvent utilisé en biologie où l'alphabet fini $\mathcal{A}$ est l'ensemble $\{\mathtt{a}, \mathtt{g}, \mathtt{c}, \mathtt{t}\}$ des 4 nucléotides, quand l'information sur la séquence est fournie par exemple par un bloc de 2 nucléotides ($m = 2$). En réécrivant alors la séquence S de longueur n en une succession de blocs de m variables aléatoires, elle devient une séquence $S^{(m)}$ de longueur $n-m+1$. Comme la séquence S est modélisée par une chaîne Mm, alors la séquence $S^{(m)}$ peut être considérée comme une séquence extraite d'une chaîne de Markov d'ordre 1 à valeurs dans l'ensemble $\mathcal{A}^m$ qui est aussi un alphabet fini.

Par analogie aux définitions données dans le modèle d'une chaîne de Markov d'ordre 1, nous donnons les définitions et les notations utilisées dans le cas de la modélisation des séquences par une chaîne de Markov d'ordre m. Nous montrons que la maximalité à gauche d'une répétition chevauchante d'un mot de longueur t dans la séquence S est identique à la maximalité à gauche d'une répétition d'un mot de longueur $t-m+1$ dans la séquence $S^{(m)}$, et que le nombre N_t de répétitions chevauchantes maximales à gauche de mots de longueur t dans séquence S, est égale au nombre $N^{(m)}_{t-m+1}$ de répétitions chevauchantes maximales à gauche de mots de longueur $t-m+1$ dans la séquence $S^{(m)}$.

Nous montrons de même, en majorant chacune des probabilités de transition de l'expression du nombre moyen $\lambda^{(m)} = \mathbb{E}\left(N^{(m)}_{t-m+1}\right)$ de répétitions chevauchantes maximales à gauche pour ce modèle par ξ $(0 < \xi < 1)$, que pour $t-m+1 = o(n)$, le nombre $\lambda^{(m)}$ est majoré par $n^2\xi^{t-m+1}$, et si de plus $n^2\xi^{t-m+1} = O(1)$, alors $\lambda^{(m)}$ est borné sur $]0, +\infty[$. Sous ces conditions nous énonçons le théorème 3.2.1 donnant l'approximation par une loi Poisson $\mathcal{P}_{\lambda^{(m)}}$ de paramètre $\lambda^{(m)}$, où $\lambda^{(m)} = \mathbb{E}\left(N^{(m)}_{t-m+1}\right)$, de la loi de probabilité $\mathcal{L}(N_t)$ de N_t, qui est en fait un corollaire du théorème 2.2.1. Nous avons donc aussi pour le modèle Mm, l'approximation de la loi de probabilité $\mathcal{L}(N_t)$ de N_t par la loi de Poisson $\mathcal{P}_{\lambda^{(m)}}$, pour de très longues séquences.

Nous terminons le manuscrit par la conclusion, où nous résumons et commentons les résultats obtenus pour le modèle markovien, et proposons les perspectives pour la suite à donner à ce travail.

Chapitre 1

Modélisation des séquences

Nous donnons dans ce chapitre la modélisation mathématique d'une séquence de longueur n et d'une répétition d'un mot de longueur t dans la séquence, ainsi que les notations qui seront utilisées dans la suite de ce travail.

Sommaire

1.1 Définitions et notations du modèle

Comme nous l'avons cité dans l'introduction, les séquences à comparer à celles observées sont générées aléatoirement. Nous supposons pour cela que $(X_n)_{n>1}$ est une suite de variables aléatoires à valeurs dans un alphabet fini $\mathcal{A}$, et que $S = X_1X_2\cdots X_n$ est une séquence de n variables aléatoires extraites de la suite $(X_n)_{n\geq 1}$.

1.1.1 Répétitions dans une séquence

Un mot de longueur t commençant à la position i (où $i = 1, \ldots, n-t+1$) est une succession de t lettres $X_i \cdots X_{i+t-1}$ dans la séquence S que nous notons par w_i.

Définition 1.1.1. *Le mot w_i à la position i est* répété *à la position j (où $j \neq i$) si et seulement si pour tout $k = 0, \ldots, t-1$, la lettre en position $i+k$ dans w_i est identique à celle en position $j+k$ dans w_j.*

On dit qu'en une position $\alpha = (i, j)$ on a une répétition *d'un mot de longueur t, si le mot à la position i est répété à la position j.*

En vertu de la définition 1.1.1, dire qu'il y a répétition en $\alpha = (i, j)$ du mot w_i revient à dire que $w_i = w_j$, ce qui est équivalent à :

$$\begin{cases} X_i = X_j \\ \vdots \\ X_{i+t-1} = X_{j+t-1} \end{cases} \qquad \text{(R)}$$

Exemple 1.1.1. Soit la séquence « S = `gatatatactactgatactac` » de longueur 21, elle contient 8 répétitions de mots de longueur 4. Le mot « `gata` » est répété en $(1, 14)$; « `atat` » en $(2, 4)$; « `tata` » en $(3, 5)$; « `atac` » en $(6, 15)$; « `tact` » en $(7, 10)$ et $(7, 16)$; « `acta` » en $(8, 17)$; « `ctac` » en $(9, 18)$ (figure 1.1),

```
1.  g a t a t a t a c t a c t g a t a c t a c
2.  g a t a t a t a c t a c t g a t a c t a c
3.  g a t a t a t a c t a c t g a t a c t a c
4.  g a t a t a t a c t a c t g a t a c t a c
5.  g a t a t a t a c t a c t g a t a c t a c
6.  g a t a t a t a c t a c t g a t a c t a c
7.  g a t a t a t a c t a c t g a t a c t a c
```

FIG. 1.1 – Les différents types de répétitions.

L'exemple 1.1.1 montre qu'il y a deux types de répétitons :
- celles où les mots répétés se recouvrent, appelées *répétitions chevauchantes*,
- celles où au moins une lettre sépare un mot de sa répétition, appelées *répétitions non chevauchantes*.

Formellement ces deux types de répétitions sont données dans la définition ci-dessous.

Définition 1.1.2. *On dit qu'à la position $\alpha = (i, j)$ une répétition est :*

1. non chevauchante *si : $i < i + t < j$;*
2. chevauchante *si : $|i - j| < t$.*

Si de plus la condition suivante (appelée condition de maximalité à gauche*) est vérifiée :*

$$\begin{cases} i = 1 \\ ou \\ X_{i-1} \neq X_{j-1} \quad si \quad i > 1 \end{cases}$$

la répétition en α est dite maximale à gauche.

Remarque 1.1.1. La condition $i = 1$ de la maximalité à gauche d'une répétition de la définition 1.1.2, signifie que l'on ne tient pas compte de la lettre qui précède la première lettre de la séquence, on dit qu'il y a *effet de bord.* La deuxième condition pour $i > 1$, signifie que la répétition ne peut être élargie à gauche.

Exemple 1.1.2. Reprenons la séquence S de l'exemple 1.1.1. Parmi les 8 répétitions de mots de longueur 4, il y a celles qui sont (figure 1.1) :

- non chevauchantes et maximales à gauche : « `gata` » en $(1, 14)$ et « `atac` » en $(6, 15)$;
- non chevauchantes et non maximales à gauche : « `tact` » en $(7, 16)$, « `acta` » en $(8, 17)$ et « `ctac` » en $(9, 18)$;
- chevauchantes et maximales à gauche : « `atat` » en $(2, 4)$ et « `tact` » en $(7, 10)$;
- chevauchantes et non maximales à gauche (il y a une seule) : « `tata` » en $(3, 5)$.

Les répétitions auxquelles nous nous intéresserons dans ce travail sont *chevauchantes et maximales à gauche*, c'est à dire celles vérifiant la condition 2. et la condition de maximalité à gauche de la définition 1.1.2.

La condition 2. de la définition 1.1.2 étant réalisée, posons :

$$\ell = |i - j| \qquad \text{où} \qquad \ell = 1, \ldots, t-1$$

Considérons par symétrie que $i < j$, la composante j est donc telle que $j = i + \ell$. La position $\alpha = (i, j)$ et la relation (R) deviennent respectivement $\alpha = (i, i + \ell)$, et :

$$w_i = w_{i+\ell} \Leftrightarrow \begin{cases} X_i = X_{i+\ell} \\ \vdots \\ X_{i+t-1} = X_{i+\ell+t-1} \end{cases}$$

ainsi la condition de maximalité à gauche devient :

$$\begin{cases} i = 1 \\ \text{ou} \\ X_{i-1} \neq X_{i+\ell-1} \quad \text{si} \quad i > 1 \end{cases} \qquad (*)$$

En une position $\alpha = (i, i + \ell)$ d'une répétition chevauchante et maximale à gauche, la position $i + \ell$ du mot répété $w_{i+\ell}$ est telle que :

$$i + 1 \leq i + \ell \leq i + t - 1$$

la répétition en $\alpha = (i, i + \ell)$ est alors un seul mot de longueur $t + \ell$.

Nous adoptons donc les natations suivantes :

- la répétition en $\alpha = (i, i + \ell)$ du mot w_i est un mot $X_i \cdots X_{i+\ell+t-1}$ de longueur $t + \ell$ que nous notons par W_α ;
- l'ensemble des positions des répétitions chevauchantes et maximales à gauche de mots de longueur t est noté :

$$\mathcal{I} = \{\alpha = (i, i + \ell) \ / \ \ell = 1, \ldots, t-1;\ i = 1, \ldots, n - t - \ell + 1\}$$

Dans toute la suite à la position α de $\mathcal{I}$, une répétition chevauchante maximale à gauche est considérée comme un mot W_α de longueur $t + \ell$.

1.1.2 Caractérisation d'une répétition chevauchante et maximale à gauche

La répétition en $\alpha = (i, i+\ell)$ d'un mot w_i de longueur t étant un mot W_α de longueur $t+\ell$, les lettres qui composent w_i ne peuvent être toutes distinctes, elles sont la répétition des ℓ premières lettres de w_i. En effet les ℓ premières lettres de w_i sont les ℓ premières de $w_{i+\ell}$, et les $t-\ell$ dernières lettres de w_i sont parmi les ℓ premières de $w_{i+\ell}$, et donc de w_i (figure 1.2),

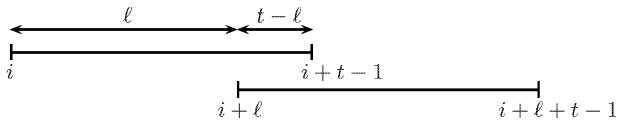

FIG. 1.2 – Les lettres de w_i ne sont pas toutes distinctes.

Le mot W_α étant lui même formé de la répétition de w_i en $\alpha = (i, i+\ell)$, il est par conséquent la répétition *périodique* des ℓ premières lettres dans cette ordre. On dit alors que ℓ est la *période* de W_α ou le mot W_α admet pour période ℓ.

Si parmi les ℓ premières lettres de w_i il y en a d (où d est le plus petit entier naturel tel que $0 < d \leq \ell$) qui sont répétées, elles sont alors répétées dans W_α, la période de W_α sera d, et c'est *la plus petite période*. Dans ce cas ℓ est un multiple de d, et la condition de maximalité à gauche $(*)$ est équivalente à :

$$\begin{cases} i = 1 \\ \text{ou} \\ X_{i-1} \neq X_{i+d-1} \quad \text{si} \quad i > 1 \end{cases} \qquad (**)$$

Exemple 1.1.3. Soit la séquence « S = `tacttgacttgacttgacttgactg` » de longueur 25. En $(2, 12)$, le mot « $\mathrm{W}_{(2,12)}$ = `acttgacttgacttgacttgact` » formé de la répétition du mot « w_2 = `acttgacttgact` » de longueur $t = 13$. Les lettres de w_2 ne sont pas toutes distinctes, elles sont composées des $\ell = 10$ premières de w_2. Le mot $\mathrm{W}_{(2,12)}$ est alors de longueur $t+\ell = 23$ ($\ell = 10$) et a pour période $d = 5$, en effet les 5 premières lettres « `acttg` » sont 2 fois répétées dans cet ordre parmi les ℓ premières lettres de w_2 (figure 1.3),

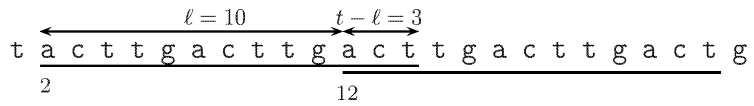

FIG. 1.3 – Les lettres de w_2 ne sont pas toutes distinctes, et $\mathrm{W}_{(2,12)}$ a pour période $d = 5$.

Dans toute la suite de ce travail nous nous plaçons dans le cadre général, où en une position α de $\mathcal{I}$ s'il y a une répétition chevauchante maximale à gauche d'un mot de longueur t, le mot W_α admet pour période d (avec $0 < d \leq \ell$).

Étant donné que le mot W_α admet pour période d $(0 < d \leq \ell)$, nous allons déterminer combien de fois des d premières sont répétées dans W_α. En effectuant la division euclidienne de $t+\ell$ par d nous obtenons $t+\ell = qd+r$ où q est le quotient et r $(0 \leq r < d)$ le reste. Ainsi les d premières lettres sont répétées q fois, et les r dernières sont les r premières des d lettres dans le même ordre, par conséquent les lettres du mot W_α sont telles que :

1. si $i = 1$:

$$\begin{cases} X_1 \cdots X_d = X_{1+d} \cdots X_{2d} = \cdots = X_{1+(q-1)d} \cdots X_{qd} \\ \text{et} \\ X_{qd+1} \cdots X_{qd+r} = X_1 \cdots X_r \end{cases} \tag{1.1}$$

2. si $i > 1$ et $X_{i-1} \neq X_{i+d-1}$:

$$\begin{cases} X_i \cdots X_{i+d-1} = X_{i+d} \cdots X_{i+2d-1} = \cdots = X_{i+(q-1)d} \cdots X_{i+qd-1} \\ \text{et} \\ X_{i+qd} \cdots X_{i+qd+r-1} = X_i \cdots X_{i+r-1} \end{cases} \tag{1.2}$$

Exemple 1.1.4. Reprenons la séquence S de l'exemple 1.1.3. D'après la division euclidienne de 23 par 5, le mot $W_{(2,12)}$ de longueur 23 est composé de la répétition de 4 ($q = 4$) fois des 5($d = 5$) lettres « `acttg` » qui forment la période, suivies des 3 ($r = 3$) lettres « `act` » (figure 1.4),

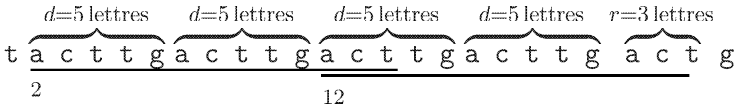

FIG. 1.4 – les lettres « `acttg` »sont répétées 4 fois, les lettres « `act` » sont les 3 restantes.

1.2 Choix du modèle markovien

La première étude faite par R.Arratia *et al* (1996) [3] est basée sur des séquences extraites d'une suite de variables aléatoires de Bernouilli indépendantes identiquement distribuées (i-i-d). Dans ce cas l'information sur la séquence est entièrement déterminée par les propriétés de chacune des lettres qui la composent indépendamment les unes des autres et donc des places qu'elles occupent dans la séquence. La probabilité qu'il y ait répétition en une position donnée dans la séquence est alors détérminée par la probabilité des occurrences des lettres qui sont des variables aléatoires indépendantes. Les cas rencontrés dans la pratique notamment en biologie (où l'alphabet $\mathcal{A}$ est l'ensemble $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$), l'information sur la séquence est fournie par 2 nucléotides voire plus, autrement dit, par des blocs de mots de longueur dépassant 1, d'où l'existence d'une dépendance entre les variables aléatoires de la séquence. Il est donc naturel de supposer la séquence aléatoire extraite d'une suite de variables aléatoires dépendantes.

La dépendance dont il s'agit dans notre travail est une dépendance markovienne, c'est à dire que les lettres d'une séquence sont extraites d'une chaîne de Markov. Nous nous intéressons plus précisément aux séquences générées par une chaîne de Markov d'ordre 1, homogène et stationnaire. Nous commençons par donner les définitions, les notations ainsi que les propriétés des chaînes de Markov pour cette modélisation.

1.2.1 Définitions et notations

Soit $(X_n)_n$ une suite de variables aléatoires dépendantes à valeur dans un alphabet fini $\mathcal{A}$.

Définition 1.2.1. *La suite $(X_n)_n$ est dite* chaîne de Markov, *si pour tout entier naturel $n \geq 1$ et tout $a_1, \ldots, a_n$ dans $\mathcal{A}$:*

$$\mathbb{P}(X_n = a_n \,/\, X_{n-1} = a_{n-1}, \ldots, X_1 = a_1) = \mathbb{P}(X_n = a_n \,/\, X_{n-1} = a_{n-1})$$

le membre de droite est appelé probabilité de transition en une étape.

L'ensemble $\mathcal{A}$ est appelé *espace des états* de la chaîne de Markov, et pour tout $i \geq 1$ et a, b de $\mathcal{A}$, la probabilité de transition et la loi initiale sont notées respectivement par :

$$\pi(a,b) = \mathbb{P}(X_n = b \,/\, X_{n-1} = a) \qquad \text{et} \qquad \mu(a) = \mathbb{P}(X_1 = a)$$

Définition 1.2.2. *Si pour tout a, b dans $\mathcal{A}$ les probabilités de transition $\pi(a,b)$ ne dépendent pas de l'instant n, la chaîne de Markov est dite* homogène *(dans le temps).*

Ce qui signifie dans notre cas que les probabilités de transition ne dépendent pas des positions des lettres dans la séquence.

Pour tout a, b dans $\mathcal{A}$ les probabilités de transition $\pi(a,b)$ forment les cœfficients d'une matrice dite *matrice de probabilités de transition*, elle est notée :

$$\Pi = (\pi(a,b))_{a,b \in \mathcal{A}}$$

Proposition 1.2.1. *Si $(X_n)_n$ est une chaîne de Markov à valeurs dans $\mathcal{A}$, alors pour tout $a_1, \ldots, a_k$ dans $\mathcal{A}$, on a :*

$$\mathbb{P}(X_1 = a_1, \ldots, X_k = a_k) = \mu(a_1) \prod_{i=1}^{k-1} \pi(a_i, a_{i+1}) \tag{1.3}$$

Une chaîne de Markov est donc caractérisée par :
- sa loi initiale μ,
- sa matrice de probabilités de transition Π.

Définition 1.2.3. *Soient $(X_n)_n$ une chaîne de Markov sur un espace d'état fini et Π sa matrice de probabilités de transition. La matrice Π est dite* stochastique *si ses cœfficients vérifient les conditions suivantes :*

$$\forall a, b \in \mathcal{A} \quad \pi(a,b) \geq 0 \qquad et \qquad \forall\, a \in \mathcal{A} \quad \sum_{b \in \mathcal{A}} \pi(a,b) = 1$$

Remarque 1.2.1. La matrice Π est de dimension
- finie si l'ensemble des états de la chaîne de Markov est fini,
- infinie si l'ensemble des états est infini dénombrable.

Dans notre cas la matrice Π est de dimension finie car $\mathcal{A}$ est un alphabet fini.

1.2.2 Chaîne de Markov stationnaire

Soit $(X_n)_n$ une chaîne de Markov à valeurs dans $\mathcal{A}$. Soient a, b deux éléments de $\mathcal{A}$, notons par :

$$\pi^{(k)}(a,b) = \mathbb{P}(X_{k+1} = b \,/\, X_1 = a)$$

la probabilité d'être en b à l'instant $k+1$ en partant de l'état a à l'instant 1.

Pour tout a, b dans $\mathcal{A}$, les nombres $\pi^{(k)}(a,b)$ représentent les cœfficients de la matrice $\Pi^{(k)}$ de probabilités de transition en k étapes.

Proposition 1.2.2. *Soit $(X_n)_n$ une chaîne de Markov à valeurs dans $\mathcal{A}$. Pour tout $k \geq 0$, la matrice de probabilités de transition en k étapes est telle que :*

$$\Pi^{(k)} = \Pi^k$$

du produit matriciel on déduit la relation suivante (équation de Chapman-Kolmogorov) :

$$\forall\, a, b \in \mathcal{A} \qquad \pi^{(k)}(a,b) = \sum_{c \in \mathcal{A}} \pi^{(k_1)}(a,c)\pi^{(k_2)}(c,b) \qquad \text{où} \qquad k_1 + k_2 = k$$

et la probabilité $\mathbb{P}(X_k = b) = \mu(b)$ d'être à l'état b à l'étape k est telle que :

$$\forall\, b \in \mathcal{A} \qquad \mathbb{P}(X_k = b) = \sum_{a \in \mathcal{A}} \mu(a)\pi^{(k-1)}(a,b) \tag{1.4}$$

Définition 1.2.4. *Soit $(X_n)_n$ une chaîne de Markov à valeurs dans $\mathcal{A}$ de loi initiale μ et de matrice de probabilités de transition Π. Si μ vérifie la relation suivante :*

$$\forall\, a \in \mathcal{A} \qquad \mu(a) = \sum_{b \in \mathcal{A}} \mu(b)\pi(b,a)$$

elle est dite invariante *pour Π, et la chaîne de Markov est dite* stationnaire.

D'après la définition 1.2.4, la relation (1.4) est vraie pour tout k. Sous forme matricielle, la relation de la définition 1.2.4 s'écrit :

$$\mathsf{P}' = \mathsf{P}'\Pi$$

où P' désigne le vecteur transposé de $\mathsf{P} = (\mu, \ldots, \mu)$ dont les composantes sont des probabilités.

Proposition 1.2.3. *Soit $(X_n)_n$ une chaîne de Markov à valeurs dans $\mathcal{A}$ de loi initiale μ et de matrice de probabilités de transition Π, si la chaîne $(X_n)_n$ est stationnaire, alors :*

$$\forall k \in \mathbb{N}^* \qquad \mathsf{P}' = \mathsf{P}'\Pi^k$$

1.3 Dénombrement des répétitions

Soit $S = X_1X_2\cdots X_n$ une séquence aléatoire de longueur n extraite d'une chaîne de Markov homogène et stationnaire $(X_n)_{n\geq 1}$ d'ordre 1, à valeurs dans un alphabet fini $\mathcal{A}$. Soit W_α le mot $X_i \cdots X_{i+t+\ell-1}$ formé en $\alpha = (i, i+\ell)$ par la répétition du mot $w_i = X_i \cdots X_{i+t-1}$ (où $i = 1, \ldots, n-t+1$; $\ell = 1, \ldots, t-1$) dans la séquence S. On note par :

$$\Pi = (\pi(a,b))_{a,b\,\in\mathcal{A}} \qquad \text{où} \qquad \forall\, a, b \in \mathcal{A} \quad \pi(a,b) > 0$$

la matrice de probabilités de transition et par μ la loi initiale qui est aussi la loi stationnaire vérifiant :

$$\forall\, a \in \mathcal{A} \qquad \mu(a) = \sum_{b \in \mathcal{A}} \mu(b)\pi(b,a)$$

1.3.1 Nombre moyen de répétitions

Soit Y_α la variable aléatoire indicatrice qui vaut 1 si une répétition occurre en une position α de $\mathcal{I}$, ou 0 sinon, elle est définie par :

$$Y_\alpha = \begin{cases} \mathbf{1}_{\{w_1 = w_{1+\ell}\}} & \text{si} \quad i = 1 \\ \mathbf{1}_{\{X_{i-1} \neq X_{i+d-1},\, w_i = w_{i+\ell}\}} & \text{si} \quad i > 1 \end{cases}$$

Pour ℓ fixé, i a $n-t-\ell+1$ positions, le nombre de répétitions N_t est donné par :

$$N_t = \sum_{\ell=1}^{t-1} \sum_{i=1}^{n-t-\ell+1} Y_\alpha$$

le nombre moyen de répétitions λ n'est autre que l'espérance mathématique $\mathbb{E}(N_t)$ du nombre de répétitions N_t. D'après la linéarité de l'espérance mathématique, λ est donné par :

$$\lambda = \sum_{\ell=1}^{t-1} \sum_{i=1}^{n-t-\ell+1} \mathbb{E}\left(Y_\alpha\right) \tag{1.5}$$

où $\mathbb{E}\left(Y_\alpha\right)$ est la probabilité d'une occurrence en α, elle est donnée par :

$$\mathbb{E}(Y_\alpha) \quad = \quad \begin{cases} \mathbb{P}(w_1 = w_{1+\ell}) & \text{si} \quad i = 1 \\ \mathbb{P}(X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}) & \text{si} \quad i > 1 \end{cases} \tag{1.6}$$

La détermination de λ revient d'après (1.5) et (1.6), à calculer en premier lieu, la probabilité de l'occurrence de la répétition en α.

Soient $a_1, \ldots, a_d$ les valeurs prises par les $t+\ell$ variables aléatoires dans (1.1) ou dans (1.2) :

• Si $i = 1$, nous avons en vertu de (1.1) :

– si $r = 0$:

$$\{w_1 = w_{1+\ell}\} = \bigcup_{a_1,\ldots,a_d \in \mathcal{A}} \left\{ \mathrm{W}_{(1,1+\ell)} = \overbrace{a_1 \ldots a_d \ldots a_1 \ldots a_d}^{q \text{ fois } a_1 \ldots a_d} \right\}$$

– si $r \neq 0$:

$$\{w_1 = w_{1+\ell}\} = \bigcup_{a_1,\ldots,a_d \in \mathcal{A}} \left\{ \mathrm{W}_{(1,1+\ell)} = \overbrace{a_1 \ldots a_d \ldots a_1 \ldots a_d}^{q \text{ fois } a_1 \ldots a_d} a_1 \ldots a_r \right\}$$

En passant au calcul de probabilités, il s'en suit en utilisant les probabilités conditionnelles et les propriétés des chaînes de Markov,

- pour $r = 0$:

$$\begin{aligned} \mathbb{P}(w_1 = w_{1+\ell}) \quad &= \quad \mathbb{P}\left(\bigcup_{a_1,\ldots,a_d \in \mathcal{A}} \left\{ \mathrm{W}_{(1,1+\ell)} = \overbrace{a_1 \ldots a_d \ldots a_1 \ldots a_d}^{q \text{ fois } a_1 \ldots a_d} \right\} \right) \\ &= \quad \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mathbb{P}\left(\mathrm{W}_{(1,1+\ell)} = \overbrace{a_1 \ldots a_d \ldots a_1 \ldots a_d}^{q \text{ fois } a_1 \ldots a_d} \right) \\ &= \quad \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mathbb{P}(X_1 = a_1, \ldots, X_d = a_d, X_{d+1} = a_1, \ldots, X_d = a_d, X_{1+d} = a_1, \\ &\qquad \ldots, X_{qd} = a_d) \\ &= \quad \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mu(a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right)^{q-1} (\pi(a_d, a_1))^{q-1} \end{aligned}$$

- pour $r \neq 0$:

$$\begin{aligned}
\mathbb{P}(w_1 = w_{1+\ell}) &= \mathbb{P}\left(\bigcup_{a_1,\ldots,a_d \in \mathcal{A}} \left\{ \mathrm{W}_{(1,1+\ell)} = \overbrace{a_1 \ldots a_d \ldots a_1 \ldots a_d}^{q \text{ fois } a_1 \ldots a_d}\, a_1 \ldots a_r \right\}\right) \\
&= \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mathbb{P}(X_1 = a_1, \ldots, X_d = a_d, X_{d+1} = a_1, \ldots, X_d = a_d, \\
&\qquad X_{1+d} = a_1, \ldots, X_{qd} = a_d, X_{qd+1} = a_1, \ldots, X_{qd+r} = a_r) \\
&= \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mu(a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1})\right) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1})\right)^{q-1} \\
&\qquad (\pi(a_d, a_1))^{q-1} \left(\pi(a_d, a_1) \prod_{j=1}^{r-1} \pi(a_j, a_{j+1})\right)
\end{aligned}$$

nous obtenons dans les deux cas :

$$\mathbb{P}(w_1 = w_{1+\ell}) = \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mu(a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1})\right) \Gamma \tag{1.7}$$

où :

$$\Gamma = \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1})\right)^{q-1} (\pi(a_d, a_1))^{q-1} \left(\pi(a_d, a_1) \prod_{j=1}^{r-1} \pi(a_j, a_{j+1}))\right)^{\mathbf{1}_{\{r \neq 0\}}} \tag{1.8}$$

• Si $i > 1$, le même raisonnement que ci-dessus, donne en vertu de (1.2) tout en tenant compte de la condition de maximalité à gauche $(**)$, pour $r = 0$:

$$\{X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}\} = \bigcup_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \left\{ X_{i-1} = b, \mathrm{W}_{(i,i+\ell)} = \overbrace{a_1 \ldots a_d \ldots a_1 \ldots a_d}^{q \text{ fois } a_1 \ldots a_d} \right\}$$

et pour $r \neq 0$:

$$\{X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}\} = \bigcup_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \left\{ X_{i-1} = b, \mathrm{W}_{(i,i+\ell)} = \overbrace{a_1 \ldots a_d \ldots a_1 \ldots a_d}^{q \text{ fois } a_1 \ldots a_d}\, a_1 \ldots a_r \right\}$$

le passage au calcul de probabilités, donne en utilisant les probabilités conditionnelles et les propriétés des chaînes de Markov, pour $r = 0$:

$$\begin{aligned}
&\mathbb{P}(X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}) \\
&= \sum_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \mathbb{P}(X_{i-1} = b, X_i = a_1, \ldots, X_{i+d-1} = a_d, X_{i+d} = a_1, \ldots, X_{i+(q-1)d-1} \\
&\qquad = a_1, \ldots, X_{i+qd-1} = a_d) \\
&= \sum_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \mu(b)\pi(b, a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1})\right) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1})\right)^{q-1} (\pi(a_d, a_1))^{q-1}
\end{aligned}$$

et pour $r \neq 0$:

$$\begin{aligned}
&\mathbb{P}(X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}) \\
&\quad = \sum_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \mathbb{P}(X_{i-1} = b, X_i = a_1, \ldots, X_{i+d-1} = a_d, X_{i+d} = a_1, \ldots, X_{i+(q-1)d-1} = a_1, \\
&\qquad \ldots, X_{i+qd-1} = a_d, X_{i+qd} = a_1, \ldots, X_{i+r-1+qd} = a_r) \\
&\quad = \sum_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \mu(b)\pi(b,a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right)^{q-1} (\pi(a_d, a_1))^{q-1} \\
&\qquad \left(\pi(a_d, a_1) \prod_{j=1}^{r-1} \pi(a_j, a_{j+1}) \right)
\end{aligned}$$

ainsi :

$$\mathbb{P}(X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}) = \sum_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \mu(b)\pi(b,a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right) \Gamma \tag{1.9}$$

où Γ est donné par (1.8).

Lemme 1.3.1. *Soient $a_1, \ldots, a_d$ les valeurs prises par les $t+\ell$ variables aléatoires dans* (1.1) *ou dans* (1.2). *Le nombre moyen de répétitions chevauchantes et maximales à gauche de mots de longueur t est :*

$$\begin{aligned}
\lambda \;=\; & \sum_{\ell=1}^{t-1} \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mu(a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right)^{q} (\pi(a_d, a_1))^{q-1} \left(\pi(a_d, a_1) \prod_{j=1}^{r-1} \pi(a_j, a_{j+1}) \right)^{\mathbf{1}_{\{r \neq 0\}}} \\
& + \sum_{\ell=1}^{t-1} (n - t - \ell) \sum_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} \mu(b)\pi(b,a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right)^{q} (\pi(a_l, a_1))^{q-1} \\
& \left(\pi(a_d, a_1) \prod_{j=1}^{r-1} \pi(a_j, a_{j+1}) \right)^{\mathbf{1}_{\{r \neq 0\}}}
\end{aligned} \tag{1.10}$$

Démonstration. Il suffit de remplacer $\mathbb{E}(Y_\alpha)$ respectivement par (1.7) pour $i = 1$ et (1.9) pour $i > 1$, dans l'expression (1.5). C.Q.F.D

1.3.2 Ordre de grandeur du nombre moyen de répétitions

Le calcul explicite de λ nécessite de calculer l'expression (1.10) et donc de connaître toutes les probabilités de transition, ce qui ne peut être déterminé avec exactitude. Néanmoins nous pouvons donner une expression qui domine λ en fonction des longueurs n et t respectivement de la séquence S et du mot w_i, et de trouver ensuite des conditions sur n et t pour que λ soit borné sur $]0, +\infty[$.

Posons :

$$\xi = \max_{a,b \in \mathcal{A}} \pi(a,b) \qquad (0 < \xi < 1)$$

Lemme 1.3.2. *La probabilité de l'occurrence de la répétition en α, est telle que :*

$$\mathbb{E}(Y_\alpha) \leq \xi^t \tag{1.11}$$

Démonstration. Soient $a_1, \ldots, a_d$ les valeurs prises par les $t+\ell$ variables aléatoires dans (1.1) ou dans (1.2). Étant donné que l'expression (1.8) est un facteur qu'on retrouve dans chacune des expressions (1.7) et (1.9), nous avons alors en majorant dans (1.8) chaque probabilité de transition par ξ $(0 < \xi < 1)$:

$$\begin{aligned} \Gamma &\leq \xi^{(q-1)(d-1)+q-1+r\mathbf{1}_{\{r\neq 0\}}} \\ &= \xi^{qd-d+r\mathbf{1}_{\{r\neq 0\}}} \\ &= \xi^{t+\ell-d} \end{aligned}$$

comme $\ell - d \geq 0$ alors $\xi^{\ell-d} \leq 1$, par suite :

$$\Gamma \leq \xi^t$$

en remplçant cette majoration dans (1.7), celle-ci devient : :

$$\mathbb{P}(w_1 = w_{1+\ell}) \leq \left(\sum_{a_1,\ldots,a_d \in \mathcal{A}} \mu(a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right) \right) \xi^t \tag{1.12}$$

et d'après les propriétés des chaînes de Markov :

$$\begin{aligned} \sum_{a_1,\ldots,a_d \in \mathcal{A}} \mu(a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right) &= \sum_{a_d \in \mathcal{A}} \left(\sum_{a_1 \in \mathcal{A}} \mu(a_1) \pi^{(d-1)}(a_1, a_d) \right) \\ &= \sum_{a_d \in \mathcal{A}} \mu(a_d) \\ &= 1 \end{aligned}$$

il s'en suit alors :

$$\mathbb{P}(w_1 = w_{1+\ell}) \leq \xi^t$$

de même nous avons en remplaçant Γ par sa majoration ξ^t dans (1.9) :

$$\mathbb{P}(X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}) \leq \left(\sum_{\substack{b,a_1,\ldots,a_l \in \mathcal{A} \\ b \neq a_d}} \mu(b)\pi(b, a_1) \left(\prod_{j=1}^{d-1} \pi(a_j, a_{j+1}) \right) \right) \xi^t$$

en utilisant les propriétés des chaînes de Markov, la somme entre parenthèses de l'expression ci-dessus est majorée par 1, et nous obtenons en remplaçant dans la même expression :

$$\mathbb{P}(X_{i-1} \neq X_{i+d-1},\ w_i = w_{i+\ell}) \leq \xi^t$$

En réunissant ces deux cas nous déduisons (1.11). C.Q.F.D

En nous basant sur le lemme 1.3.1, nous allons déterminer des conditions sur n et t pour que λ soit borné sur $]0, +\infty[$.

Proposition 1.3.1. *Pour $t = o(n)$ on a :*

$$\lambda \leq n^2 \xi^t$$

Si $n^2\xi^t = O(1)$, alors λ est borné sur $]0, +\infty[$.

Démonstration. En appliquant le lemme 1.3.2 pour majorer l'expression (1.10) du lemme 1.3.1, nous aurons :

$$\begin{aligned} \lambda &\leq \sum_{\ell=1}^{t-1} \xi^t + \sum_{\ell=1}^{t-1} (n - t - \ell)\, \xi^t \\ &= \left((t-1) + (t-1)(n-t) - \sum_{\ell=1}^{t-1} \ell \right) \xi^t \\ &= (t-1)\left(1 + n - t - \frac{t}{2}\right) \\ &= \left(1 + \frac{1}{n} - \frac{3}{2}\frac{t}{n}\right) n(t-1)\, \xi^t \end{aligned} \tag{1.13}$$

Pour $t = o(n)$ et $t \geq 1$, nous avons :

$$1 + \frac{1}{n} - \frac{3}{2}\frac{t}{n} \leq 1$$

après avoir majoré $t - 1$ par n, nous obtenons en remplaçant dans (1.13) :

$$\lambda \leq n^2 \xi^t$$

Si de plus $n^2\xi^t$ est borné par une constante positive, c'est à dire $n^2\xi^t = O(1)$, alors λ est borné sur $]0, +\infty[$. C.Q.F.D

Corollaire 1.3.1. *Si $t = O\left(\log_{\frac{1}{\xi}}(n)\right)$, alors λ est borné sur $]0, +\infty[$.*

Démonstration. Il s'agit de trouver l'ordre de grandeur de t par rapport à n. Si $n^2\xi^t = O(1)$, alors :

$$\begin{aligned} n^2\xi^t \asymp 1 &\Leftrightarrow \xi^t \asymp \frac{1}{n^2} \\ &\Leftrightarrow t \ln(\xi) \asymp 2 \ln\left(\frac{1}{n}\right) \\ &\Leftrightarrow t \asymp 2 \log_{\frac{1}{\xi}}(n) \end{aligned}$$

par conséquent :

$$n^2\xi^t = O(1) \Leftrightarrow t = O\left(\log_{\frac{1}{\xi}}(n)\right)$$

ce qui donne en appliquant la proposition 1.3.1, que λ est borné sur $]0, +\infty[$. C.Q.F.D

Chapitre 2

Approximation par la méthode de Chen-Stein

Dans le cas où la séquence aléatoire est extraite d'une suite de variables aléatoires de Bernouilli indépendantes identiquement distribuées (i-i-d), R.Arratia and *al.* (1996) [3] ont montré en utilisant la méthode de Chen-Stein, que la loi de probabilité du nombre de répétitions maximales à gauche est approximée par une loi de Poisson. Dans notre cas l'étude de l'approximation de la loi de probabilité de N_t repose aussi sur la méthode de Chen-Stein, en nous basant sur celle-ci nous énonçons le théorème donnant l'approximation de la loi de probabilité de N_t par une loi de Poisson pour notre modèle. Ce qui nous permet de calculer ou du moins de donner une approximation de la p-value.

Sommaire

2.1 Méthode de Chen-Stein et choix du voisinage

2.1.1 Méthode de Chen-Stein

La méthode de Chen-Stein consiste à déterminer la borne supérieure de la *distance en variation totale* notée d_{VT} entre deux lois de probabilité, elle est appelée *borne de Chen-Stein.*

Définition 2.1.1. *Soient* $(\Omega, \mathcal{F}, \mathbb{P})$ *un espace probabilisé, soient* X *et* Y *deux variables aléatoires de lois de probabilités respectives* $\mathcal{L}(X)$ *et* $\mathcal{L}(Y)$*, la* distance en variation totale *entre* $\mathcal{L}(X)$ *et* $\mathcal{L}(Y)$ *est définie par :*

$$d_{VT}\left(\mathcal{L}(X), \mathcal{L}(Y)\right) = \sup_{\mathrm{A} \in \mathcal{F}} |\mathbb{P}(X \in \mathrm{A}) - \mathbb{P}(Y \in \mathrm{A})|$$

Pour notre modèle, il s'agit donc de déterminer la borne supérieure $d_{VT}(\mathcal{L}(N_t), \mathcal{P}_\lambda)$ où $\mathcal{L}(N_t)$ désigne la loi de N_t et $\mathcal{P}_\lambda$ la loi de Poisson de paramètre λ, où $\lambda = \mathbb{E}(N_t)$ est donné par (1.10).

R.Arratia *and al* (1990) [2] ont montré que la borne de Chen-Stein fait intervenir trois quantités b_1, b_2 et b_3, c'est à dire que la majoration de la distance en variation totale est telle que :

$$d_{VT}(\mathcal{L}(N_t), \mathcal{P}_\lambda) \leq 2\left(b_1 + b_2 + b_3\right)$$

où les quantités b_1, b_2 et b_3 sont définies respectivement par :

$$b_1 = \sum_{\alpha \in \mathcal{I}} \sum_{\beta \in \mathrm{B}_\alpha} \mathbb{E}\left(Y_\alpha\right) \mathbb{E}\left(Y_\beta\right)$$

$$b_2 = \sum_{\alpha \in \mathcal{I}} \sum_{\substack{\beta \in \mathrm{B}_\alpha \\ \beta \neq \alpha}} \mathbb{E}\left(Y_\alpha Y_\beta\right)$$

$$b_3 = \sum_{\alpha \in \mathcal{I}} \mathbb{E}\left|\mathbb{E}\left(Y_\alpha - \mathbb{E}\left(Y_\alpha\right) | \sigma\left(Y_\beta;\ \beta \notin \mathrm{B}_\alpha\right)\right)\right|$$

B_α étant le voisinage de α, c'est à dire est une partie de $\mathcal{I}$ contenant les indices β *voisins* (nous définirons cette notion dans la sous-section 2.1.2 ci-dessous) de α.

En nous basant sur l'expression ci-dessus de la borne de Chen-Stein établie par R.Arratia *and al* (1990) [2], nous allons montrer pour notre modèle que sous les conditions de la proposition 1.3.1 et du corollaire 1.3.1 (chapitre 1), la distance en variation totale $d_{VT}(\mathcal{L}(N_t), \mathcal{P}_\lambda)$ converge vers 0, et donc $\mathcal{L}(N_t)$ est approximée par la loi de Poisson $\mathcal{P}_\lambda$.

2.1.2 Choix du voisinage

Les quantités b_1, b_2 et b_3 dépendent de B_α, il est donc indispensable de bien choisir le voisinage B_α. Dans le cas où la séquence aléatoire est extraite d'une suite de variables aléatoires indépendantes identiquement distribuées (i-i-d), le voisinage B_α est choisi en général de telle sorte que si β n'appartient pas à B_α, la variable Y_α est indépendante de Y_β ($\beta \notin \mathrm{B}_\alpha$), et donc $\mathbb{E}\left(\mathbb{E}\left(Y_\alpha\right) | \sigma\left(Y_\beta;\ \beta \notin \mathrm{B}_\alpha\right)\right) = \mathbb{E}\left(Y_\alpha\right)$, par conséquent $b_3 = 0$ (R.Arratia and *and al.* (1996)[3]).

Dans notre cas la séquence aléatoire est extraite d'une chaîne de Markov, si β n'appartient pas à B_α alors les variables Y_β et Y_α ne sont pas indépendantes, ainsi $b_3 \neq 0$, d'où la nécessité de calculer ou du moins de majorer b_3.

Soient $\alpha = (i, i + \ell)$ et $\beta = (i', i' + \ell')$ où $\ell, \ell' = 1, \ldots, t - 1$ deux éléments de $\mathcal{I}$ (positions de deux répétitions de mots de longueur t dans la séquence S de longueur n).

Définition 2.1.2. *On dit que les positions* α *et* β *sont :*

1. non voisines *s'il existe au moins une lettre qui sépare les mots* W_α *et* W_β,
2. voisines *si elles ne sont pas non voisines.*

L'ensemble des éléments β *de* $\mathcal{I}$ *voisins de* α *est appelé* voisinage *de* α*, il est noté* B_α.

D'après la définition 2.1.2, les composantes des positions $\alpha = (i, i+\ell)$ et $\beta = (i', i'+\ell')$ non voisines sont telles que :

$$i' - i - \ell > t \qquad \text{ou} \qquad i - i' - \ell' > t$$

suivant le cas où le mot W_α précède ou non le mot W_β (figure 2.1, le cas où le mot W_α précède le mot W_β).

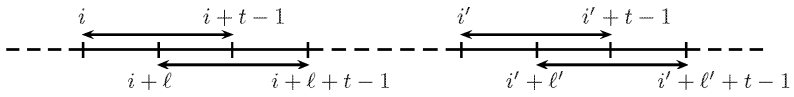

FIG. 2.1 – Les positions $\alpha = (i, i+\ell)$ et $\beta = (i', i'+\ell')$ où $\ell, \ell' = 1, \ldots, t-1$ sont non voisines.

Il vient aussi de la définition 2.1.2 que pour tout α, β dans B_α, les mots W_α et W_β forment un seul mot $\mathcal{W}_{\alpha,\beta}$ dont la longueur $|\mathcal{W}_{\alpha,\beta}|$ est telle que :

$$|\mathcal{W}_{\alpha,\beta}|_{\min} \leq |\mathcal{W}_{\alpha,\beta}| \leq |\mathcal{W}_{\alpha,\beta}|_{\max}$$

où $|\mathcal{W}_{\alpha,\beta}|_{\min}$ et $|\mathcal{W}_{\alpha,\beta}|_{\max}$ désignent respectivement les longueurs minimale et maximale du mot $\mathcal{W}_{\alpha,\beta}$, elles sont données par :

$$|\mathcal{W}_{\alpha,\beta}|_{\min} = t + \max(\ell, \ell') \qquad \text{et} \qquad |\mathcal{W}_{\alpha,\beta}|_{\max} = 2t + \ell + \ell'$$

2.2 Théorème d'approximation

2.2.1 Conditions suffisantes de maximalité à gauche

Les répétitions auxquelles nous nous intéressons étant chevauchantes et maximales à gauche, d'après la définition 2.1.2, il existe (sauf un seul cas) un recouvrement entre les mots W_α et W_β c'est à dire des recouvrements entre les mots qui forment les répétitions en α et en β. Ces recouvrements conduisent à des situations où la répétition en β (si celle-ci est précédée par celle en α) ou la répétition en α (dans le cas où elle est précédée par celle en β) n'est pas forcément maximale à gauche. Afin de ne considérer que les répétitions maximales à gauche, nous sommes donc ramené à donner des conditions suffisantes assurant la maximalité à gauche d'une répétition lorsque de telles situations se présentent.

Supposons par symétrie que la répétition en α précède celle en β, et est maximale à gauche. D'après la définition 2.1.2, les mots W_α et W_β dans cet ordre forment le mot $\mathcal{W}_{\alpha,\beta}$. La proposition et les corollaires qui vont suivre, donnent dans tous les cas de recouvrement, des conditions suffisantes pour que la répétition en β soit maximale à gauche.

Proposition 2.2.1. *Soient $\alpha = (i, i+\ell)$ et $\beta = (i', i'+\ell')$ deux positions voisines de W_α et W_β, et soient $d\,(d \leq \ell)$ et $d'\,(d' \leq \ell')$ leurs périodes respectives. Si les d et les d' lettres appartiennent aux lettres communes à W_α et W_β, alors la répétition en β n'est pas maximale à gauche s'il existe des entiers naturels non nuls K et K' tels que les lettres $X_{i'-1+Kd}$ et $X_{i'-1+K'd'}$ sont à la même position dans le mot $\mathcal{W}_{\alpha,\beta}$.*

Démonstration. Soit $X_{i'}$ la première des lettres communes (elle appartient à W_β), alors la lettre $X_{i'-1}$ appartient à W_α. Comme d et d' sont les périodes de W_α et W_β, alors il existe K et K'_1 dans $\mathbb{N}^*$ tels que $X_{i'-1+Kd}$ et $X_{i'-1+\ell'-K'_1d'}$ soient des lettres communes à W_α et W_β, il s'en suit que :

$$X_{i'-1} = X_{i'-1+Kd} \qquad \text{et} \qquad X_{i'+\ell'-1} = X_{i'-1+\ell'-K'_1d'} \tag{2.1}$$

nous déduisons aussi du fait que d' est la période de W_β, que :

$$\exists K'_2 \in \mathbb{N}^* \, (K'_2 > K'_1) \; / \; \ell' = K'_2 d'$$

donc :

$$X_{i'+\ell'-1} = X_{i'-1+(K'_2-K'_1)d'} \tag{2.2}$$

si $X_{i'-1+Kd}$ et $X_{i'-1+K'd'}$ (où $K' = K'_2 - K'_1$) sont à la même position, elles représentent la même lettre, ce qui donne d'après (2.1) et (2.2) que :

$$X_{i'-1} = X_{i'+d-1}$$

la répétition en β n'est donc pas maximale à gauche. C.Q.F.D

Corollaire 2.2.1. *Si les d et d' lettres W_α et W_β sont répétées parmi les lettres communes à W_α et W_β, alors :*

$$\ell \wedge \ell' = d = d'$$

est la période de $\mathcal{W}_{\alpha,\beta}$, et la répétition en β n'est pas maximale à gauche.

Démonstration. Supposons par symétrie que $d \leq d'$. Nous avons alors :

$$X_{i'} = X_{i'+d} = X_{i'+d'} = X_{i'+(d'-d)}$$

Si $d' - d > d$ ou $d' - d < d$ nous aurons aussi :

$$X_{i'} = X_{i'+(d'-2d)} \qquad \text{ou} \qquad X_{i'} = X_{i'+(2d-d')}$$

en continuant le même procédé, nous obtenons en utilisant l'algorithme d'Euclide que :

$$X_{i'} = X_{i'+(d\wedge d')}$$

où $d \wedge d'$ désigne le p.g.c.d de d et d', qui est aussi le p.g.c.d de ℓ et ℓ' (car ℓ est un multiple de d, et ℓ' est un multiple de d'). Il s'en suit alors que :

$$d \wedge d' = \ell \wedge \ell'$$

est la période de W_α et W_β et donc de $\mathcal{W}_{\alpha,\beta}$. Comme d et d' sont les plus petites périodes de W_α et W_β respectivement, alors :

$$d = d'$$

et en prenant $K = K' = 1$ dans la proposition 2.2.1, nous déduisons que :

$$X_{i'-1} = X_{i'+d-1}$$

ainsi, la répétition en β n'est pas maximale à gauche. C.Q.F.D

Corollaire 2.2.2. *Si les composantes des positions $\alpha = (i, i+\ell)$ et $\beta = (i', i'+\ell')$ des deux répétitions sont telles que :*

$$i < i' < i' + \ell' \leq i + t$$

alors la répétition en β n'est pas maximale à gauche.

Démonstration. D'après les hypothèses la composante $i' + \ell'$ de la répétition en β (c'est à dire la position du mot $w'_{i'+\ell'}$) est telle que :

$$i' + \ell' \leq i + t$$

alors les ℓ' (donc les d') premières de lettres de W_β appartiennent à celles du mot en i de la répétition en α (c'est à dire à w_i), elles sont donc répétées parmi les lettres communes à W_α et W_β, ce qui donne en appliquant le corollaire 2.2.2 que :

$$\ell \wedge \ell' = d \wedge d'$$

est la période de $\mathcal{W}_{\alpha,\beta}$, ainsi la répétition en β n'est pas maximale à gauche. C.Q.F.D

2.2.2 Énoncé du théorème d'approximation

Nous nous plaçons dans les hypothèses de la proposition 1.3.1 et du corollaire 1.3.1 (chapitre 1), sous lesquelles le nombre moyen de répétitions chevauchantes maximales à gauche λ est borné sur $]0, +\infty[$. Le théorème qui va suivre donne l'approximation par la loi de Poisson de la loi de probabilité de N_t pour le modèle considéré.

Théorème 2.2.1. *Soit S une séquence de longueur n générée par une chaine de Markov homogène et stationnaire, d'ordre 1 à valeurs sur un alphabet fini $\mathcal{A}$. Soient N_t le nombre de répétitions chevauchantes maximales à gauche de mots de longueur t dans S, et $\lambda = \mathbb{E}(N_t)$ le nombre moyen de répétitions donné dans* (1.10)*. Si $t = o(n)$, il existe deux fonctions positives φ et ψ definies par :*

$$\varphi(t,n) = \frac{24t^3 + t}{n^3} + \frac{3t^2}{n^2} \qquad \textit{and} \qquad \psi(t,n) = \frac{t + 4C(\xi)}{n}$$

où $C(\xi)$ est une constante positive telle que :

$$d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right) \leq 2\left(\varphi(t,n)\left(n^2\xi^t\right)^2 + \psi(t,n)\, n^2\xi^t\right)$$

Si de plus $n^2\xi^t = O(1)$, alors :

$$\lim_{n\to+\infty} d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right) = 0$$

La démonstration du théorème 2.2.1, nécessite d'après la borne de Chen-Stein exprimée par R.Arratia *and al* (1990) [2], de tenir compte de tous les cas de recouvrement des répétitions en α et β, elle fera l'objet de la section qui va suivre.

2.3 Démonstration du théorème d'approximation

Comme nous l'avons cité dans la sous-section 2.1.1, la borne de Chen-Stein fait intervenir les quantités b_1, b_2 et b_3, cela revient à montrer plus précisément que sous les conditions du théorème 2.2.1, la somme de b_1, b_2 et b_3 est telle que :

$$b_1 + b_2 + b_3 \leq \varphi(t,n)\left(n^2\xi^t\right)^2 + \psi(t,n)\, n^2\xi^t$$

il suffit donc pour cela de majorer chacune des quantités b_1, b_2 et b_3.

2.3.1 Majoration de b_1

Soient $\alpha = (i, i+\ell)$ et $\beta = (i', i'+\ell')$ deux éléments de $\mathcal{I}$ tels que pour tout α fixé, β appartient au voisinage B_α.

Proposition 2.3.1. *Pour $t = o(n)$, nous avons :*

$$b_1 \leq 3\left(\frac{t}{n}\right)^2 (n^2\xi^t)^2$$

Démonstration. Nous avons d'après l'expression de la quantité b_1 :

$$\begin{aligned} b_1 &= \sum_{\alpha\in\mathcal{I}} \sum_{\beta\in \mathrm{B}_\alpha} \mathbb{E}(Y_\alpha)\mathbb{E}(Y_\beta) \\ &= \sum_{\alpha\in\mathcal{I}} \mathbb{E}(Y_\alpha) \left(\sum_{\beta\in \mathrm{B}_\alpha} \mathbb{E}(Y_\beta) \right) \end{aligned}$$

et en remplaçant $\mathbb{E}(Y_\beta)$ par sa majoration dans (1.11), nous obtenons :

$$b_1 \leq \sum_{\alpha\in\mathcal{I}} \mathbb{E}(Y_\alpha)|\mathrm{B}_\alpha|\xi^t \tag{2.3}$$

où $|\mathrm{B}_\alpha|$ est le nombre de positions β voisines de α, ainsi majorer b_1 revient à majorer $|\mathrm{B}_\alpha|$.

La position α de la répétition en α étant fixée, les positions de β de la répétition en β sont déterminées suivant que celle-ci occurre avant ou après celle en α. Pour cela deux cas sont à considérer.

1. **La répétition en β occurre avant celle en α :**
   La composante i' de β est alors telle que $i' < i$ (figure 2.2),

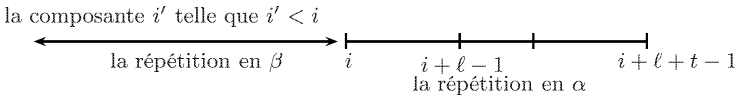

FIG. 2.2 – La répétition en β occurre avant celle en α.

   pour chaque valeur de ℓ' fixé, il y a $t+\ell'$ positions possibles pour i' et donc $(t+\ell')\ell'$ pour la position β, et comme $\ell' = 1, \ldots, t-1$ alors le nombre de positions possibles pour la répétition en β est :

$$n_{1,b_1} = (2t-1)(t-1)$$

2. **La répétition en β occurre après celle en α :**
   La composante i' de β est dans ce cas telle que $i \leq i'$. Là aussi deux cas sont à envisager, celui où i' appartient à $\{i, \ldots, i+\ell+t-1\}$, et celui où $i' > i+\ell+t-1$ (figure 2.3),

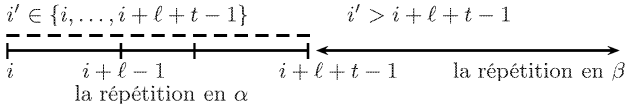

FIG. 2.3 – La répétition en β occure après celle en α.

(a) pour tout i' dans $\{i, \ldots, i+\ell+t-1\}$, il y a pour ℓ fixé $t+\ell$ positions possibles pour i' et donc $(t+\ell)\ell$ positions possibles de β, or $\ell = 1, \ldots, t-1$, alors le nombre de positions de β est $(2t-1)(t-1)$. Comme d'après le corollaire 2.2.1 et le corollaire 2.2.2, parmi ces positions il y en a au plus $(t-1)^2$ pour lesquelles la répétition en β n'est pas maximale à gauche, ainsi le nombre positions possibles de β dans ce cas est :

$$\begin{aligned} n_{2,b_1} &= (2t-1)(t-1) - (t-1)^2 \\ &= t(t-1) \end{aligned}$$

(b) pour $i' > i+\ell+t-1$, il y a d'après la définition 2.1.2 une seule position pour i', c'est celle qui correspond à $i' = i+\ell+t$, et comme $\ell' = 1, \ldots, t-1$, alors le nombre de positions possibles pour la répétition en β dans ce cas est :

$$n_{3,b_1} = t-1$$

Il en découle de 1. et 2. que :

$$\begin{aligned} |\mathrm{B}_\alpha| &\leq n_{1,b_1} + n_{2,b_1} + n_{3,b_1} \\ &= (2t-1)(t-1) + t(t-1) + (t-1) \\ &= 3t(t-1) \end{aligned}$$

ce qui nous donne après avoir majoré $t-1$ par t :

$$|\mathrm{B}_\alpha| \leq 3t^2$$

en remplaçant dans (2.3), nous obtenons :

$$\begin{aligned} b_1 &\leq 3t^2 \left(\sum_{\alpha \in \mathcal{I}} \mathbb{E}(Y_\alpha) \right) \xi^t \\ &= 3t^2 \lambda \xi^t \end{aligned} \tag{2.4}$$

et en appliquant la proposition 1.3.1, nous déduisons en remplaçant λ par sa majoration :

$$\begin{aligned} b_1 &\leq 3t^2 n^2 \xi^{2t} \\ &= 3 \left(\frac{t}{n} \right)^2 \left(n^2 \xi^t \right)^2 \end{aligned}$$

C.Q.F.D

2.3.2 Majoration de b_2

D'après l'expression de b_2, les positions $\alpha = (i, i+\ell)$ et $\beta = (i', i'+\ell')$ des répétitions en α et β sont telles que :

$$(\alpha, \beta) \in \mathcal{I}^2 \ / \ \beta \in \mathrm{B}_\alpha \quad \text{et} \quad \beta \neq \alpha \qquad (*)$$

Supposons que la répétition en α précède celle en β (le même résultat est obtenu si la répétition en α est précédée par celle en β).

Il vient de la définition 2.1.2 que la longueur $|\mathcal{W}_{\alpha,\beta}|$ de $\mathcal{W}_{\alpha,\beta}$ est déterminée suivant qu'un mot ou les deux de la répétition en β recouvrent ou non la répétition en α. Étant donné que $Y_\alpha Y_\beta$ détermine la réalisation ou non du mot $\mathcal{W}_{\alpha,\beta}$, calculer $\mathbb{E}(Y_\alpha Y_\beta)$ revient donc à calculer la probabilité que le mot $\mathcal{W}_{\alpha,\beta}$ occurre à la position (α,β). Cette probabilité dépend de la longueur $|\mathcal{W}_{\alpha,\beta}|$ du mot $\mathcal{W}_{\alpha,\beta}$, c'est à dire du nombre de lettres qui forment $\mathcal{W}_{\alpha,\beta}$, et donc du recouvrement qu'il y a entre les mots W_α et W_β.

En vertu de la définition 2.1.2, nous avons donc neuf cas de recouvrement entre les deux répétitions. Le premier cas que nous notons $k = 0$, est celui où aucune lettre ne sépare les mots W_α et W_β (figure 2.4).

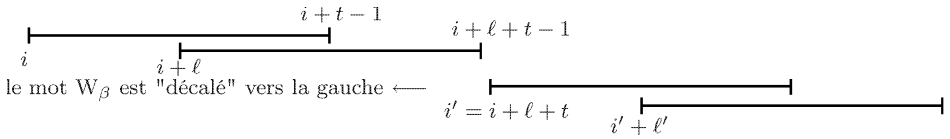

FIG. 2.4 – $k = 0$: Cas aucune lettre ne sépare W_α et W_β.

les huit autres cas $k = 1,\ldots,8$ sont obtenus en "décalant" la répétition en β vers la gauche, autrement dit, ce sont les cas où il y a réellement des recouvrements de la répétition en α par au moins un mot de la répétition en β (figure 2.5,...,2.12).

Soit G_k $(k = 0,\ldots,8)$ l'ensemble des couples (α,β) de $\mathcal{I}^2$ vérifiant $(*)$, l'expression de b_2 est réécrite comme suit :

$$b_2 = 2\left(\sum_{k=1}^{8} b_{2,k}\right) \tag{2.5}$$

où :

$$b_{2,k} = \sum_{\alpha,\beta\in \mathrm{G}_k} \mathbb{E}\left(Y_\alpha Y_\beta\right) \tag{2.6}$$

Comme $Y_\alpha Y_\beta$ est un produit de deux fonctions indicatrices, alors majorer $b_{2,k}$ revient à majorer la probabilité de l'occurrence du mot $\mathcal{W}_{\alpha,\beta}$. Celle-ci dépend comme nous l'avons dit ci-dessus de la longueur $|\mathcal{W}_{\alpha,\beta}|$ de $\mathcal{W}_{\alpha,\beta}$, ce qui nous ramène à répartir les neuf cas en deux principaux suivant la longueur $|\mathcal{W}_{\alpha,\beta}|$ de $\mathcal{W}_{\alpha,\beta}$:

- celui où $|\mathcal{W}_{\alpha,\beta}| \geq 2t$ pour $k = 0,\ldots,5$ (figures 2.5,...,2.9),
- celui où $|\mathcal{W}_{\alpha,\beta}| < 2t$ pour $k = 6,7,8$ (figures 2.10, 2.11 et 2.12).

Remarquons que le cas $|\mathcal{W}_{\alpha,\beta}| = 2t$, réalisé pour $k = 4$ ou $k = 5$ (figures 2.8 et 2.9), où la position $i' + \ell'$ du mot $w'_{i'+\ell'}$ de la répétition en β est telle que $i' + \ell' = i + t$ avec $i + \ell \leq i'$ ou $i + \ell \geq i'$, dans ce cas la répétition en β n'est pas maximale à gauche. En effet :

1. si $i + \ell \leq i'$, les deux répétitions ont $\ell + \ell'$ lettres communes (figure 2.13),

   nous avons ainsi :

   (a) si $d < \ell$ et $d' < \ell'$, les d et d' premières lettres des répétitions en α et β sont répétées parmi les $\ell+\ell'$ lettres communes aux deux répétitions, ce qui rend d'après le corollaire 2.2.1, la répétition en β non maximale à gauche.

   (b) si $d = \ell$ et $d' = \ell'$, les lettres $X_{i'-1}$, $X_{i'+\ell'-1}$ et $X_{i'+\ell+\ell'-1}$ appartiennent à la répétition en α et sont identiques, et donc la répétition en β n'est pas maximale à gauche.

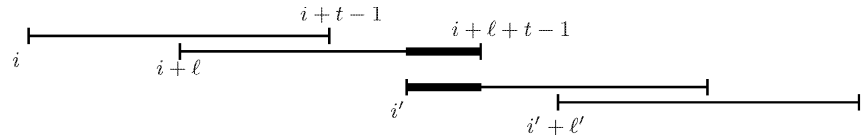

FIG. 2.5 – $k = 1 : i < i + \ell \leq i + t - 1 < i' \leq i + \ell + t - 1 < i' + \ell'$.

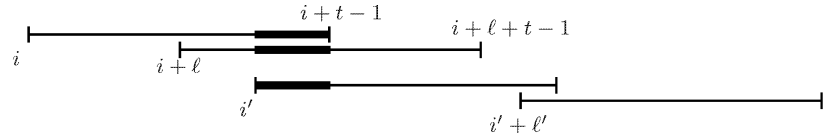

FIG. 2.6 – $k = 2 : i < i + \ell < i' \leq i + t - 1 < i + \ell + t - 1 < i' + \ell'$.

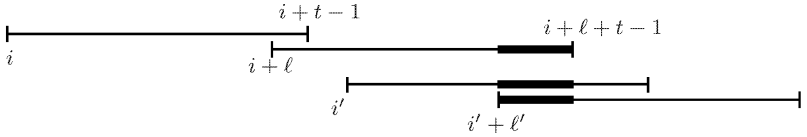

FIG. 2.7 – $k = 3 : i < i + \ell \leq i + t - 1 < i' < i + \ell \leq i' + \ell' + t - 1$.

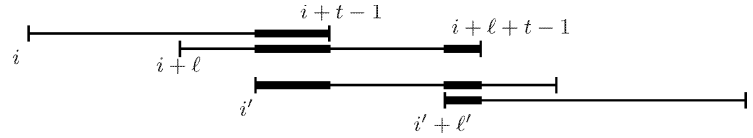

FIG. 2.8 – $k = 4 : i < i + \ell \leq i' \leq i + t - 1 < i' + \ell' \leq i + \ell + t - 1$.

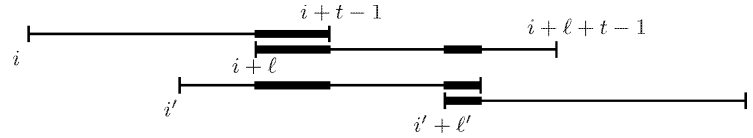

FIG. 2.9 – $k = 5 : i < i' \leq i + \ell \leq i + t - 1 < i' + \ell' \leq i' + t - 1$.

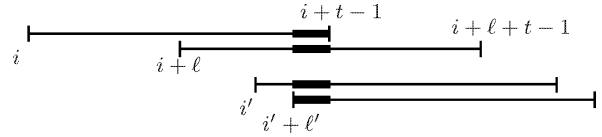

FIG. 2.10 – $k = 6 : i < i + \ell \leq i' < i' + \ell' \leq i + t - 1$.

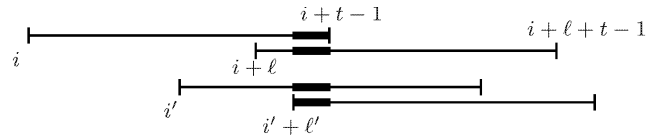

FIG. 2.11 – $k = 7 : i \leq i' < i + \ell \leq i' + \ell' \leq i + t - 1$ et $(i, i + \ell) \neq (i', i' + \ell')$.

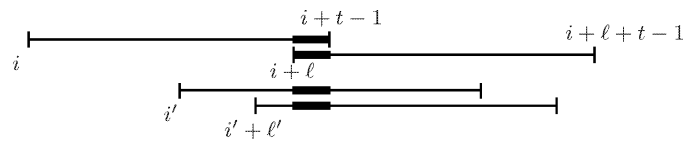

FIG. 2.12 – $k = 8 : i \leq i' < i' + \ell' \leq i + \ell \leq i + t - 1$ et $(i, i + \ell) \neq (i', i' + \ell')$.

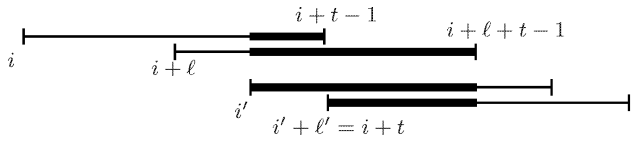

FIG. 2.13 – $|\mathcal{W}_{\alpha,\beta}| = 2t$ avec $i < i+\ell \leq i' \leq i+t-1 < i'+\ell' = i+t$.

2. si $i+\ell \geq i'$, les lettres communes aux deux répétitions sont les t lettres du mot $w_{i+\ell}$ de la répétition en α (figure 2.14),

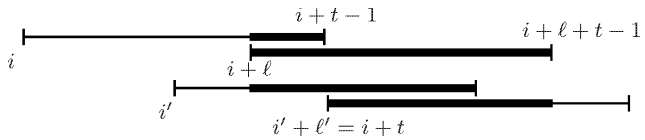

FIG. 2.14 – $|\mathcal{W}_{\alpha,\beta}| = 2t$ avec $i < i' \leq i+\ell \leq i+t-1 < i'+\ell' = i+t$.

et les d ($d \leq \ell$) et d' ($d' \leq \ell'$) lettres des périodes des répétitions en α et β sont répétées parmi ces t lettres communes, et nous obtenons là aussi d'après le corollaire 2.2.1, que la répétition en β n'est pas maximale à gauche.

Il en résulte de 1. et 2. que si $|\mathcal{W}_{\alpha,\beta}| = 2t$, la répétition en β n'est pas maximale à gauche, ainsi le cas $|\mathcal{W}_{\alpha,\beta}| \geq 2t$ est réduit au cas $|\mathcal{W}_{\alpha,\beta}| > 2t$.

La relation (2.6) se décompose donc comme suit :

$$b_{2,k} = b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} + b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} \tag{2.7}$$

où :

$$b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} = \sum_{\alpha,\beta\in \mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}}} \mathbb{E}\left(Y_\alpha Y_\beta\right) \tag{2.8}$$

et

$$b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \sum_{\alpha,\beta\in \mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}} \mathbb{E}\left(Y_\alpha Y_\beta\right) \tag{2.9}$$

où :

$$\mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} = \{(\alpha,\beta) \in \mathrm{G}_k \ / \ |\mathcal{W}_{\alpha,\beta}| > 2t\} \ \text{ et } \ \mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \{(\alpha,\beta) \in \mathrm{G}_k \ / \ |\mathcal{W}_{\alpha,\beta}| < 2t\}$$

Comme les six premiers cas ($k = 0, \ldots, 5$) correspondent à $|\mathcal{W}_{\alpha,\beta}| > 2t$, et les trois derniers ($k = 6, 7, 8$) correspondent à $|\mathcal{W}_{\alpha,\beta}| < 2t$, en posant :

$$b_{2,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} = \sum_{k=0}^{5} b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} \tag{2.10}$$

et

$$b_{2,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \sum_{k=6}^{8} b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} \tag{2.11}$$

la relation (2.5) est équivalente à :

$$b_2 = 2\left(b_{2,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} + b_{2,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}\right) \tag{2.12}$$

majorer b_2 revient donc à majorer $b_{2,\{|\mathcal{W}_{\alpha,\beta}|>2t\}}$ et $b_{2,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$

1^er cas : la longueur $|\mathcal{W}_{\alpha,\beta}|$ du mot $\mathcal{W}_{\alpha,\beta}$ est telle que $|\mathcal{W}_{\alpha,\beta}| > 2t$.

Lemme 2.3.1. *Si la longueur $|\mathcal{W}_{\alpha,\beta}|$ du mot $\mathcal{W}_{\alpha,\beta}$ est telle que $|\mathcal{W}_{\alpha,\beta}| > 2t$, alors pour tout $k = 0, \ldots, 5$:*

$$|\mathcal{W}_{\alpha,\beta}| - n_{k,b_2} \geq 2t \qquad et \qquad \mathbb{E}(Y_\alpha Y_\beta) \leq \xi^{2t}$$

où n_{k,b_2} est le nombre de lettres a priori différentes répétées dans $\mathcal{W}_{\alpha,\beta}$, et on a :

$$b_{2,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} \leq 12 \left(\frac{t}{n}\right)^3 \left(n^2 \xi^t\right)^2 \tag{2.13}$$

Démonstration. D'après l'expression (2.10), et donc (2.8), il suffit de montrer en un premier temps que pour tout $k = 0, \ldots, 5$:

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) \leq \xi^{2t} \tag{2.14}$$

Étant donné que le calcul de $\mathbb{E}\left(Y_\alpha Y_\beta\right)$ dépend du nombre de lettres qui forment le mot $\mathcal{W}_{\alpha,\beta}$, et donc de sa longueur, nous procédons en deux étapes :

- la première étape, est celle où la longueur de $\mathcal{W}_{\alpha,\beta}$ est maximale (c'est à dire $|\mathcal{W}_{\alpha,\beta}| = |\mathcal{W}_{\alpha,\beta}|_{\max}$) ;
- la deuxième étape, est celle où la longueur de $\mathcal{W}_{\alpha,\beta}$ n'est pas maximale (c'est à dire $2t < |\mathcal{W}_{\alpha,\beta}| < |\mathcal{W}_{\alpha,\beta}|_{\max}$).

Nous commençons par traiter le cas où la longueur du mot $\mathcal{W}_{\alpha,\beta}$ est maximale (figure 2.4), du fait que $Y_\alpha Y_\beta$ s'écrit de manière explicite, ce qui nous permet d'exprimer $\mathbb{E}\left(Y_\alpha Y_\beta\right)$ en fonction de toutes les probabilités de transition d'une lettre à une autre du mot $\mathcal{W}_{\alpha,\beta}$.

A- Cas où la longueur $|\mathcal{W}_{\alpha,\beta}|$ est maximale.

Ce cas correspond à celui où aucune lettre ne sépare W_α et W_β (figure 2.4), la longueur de $\mathcal{W}_{\alpha,\beta}$ est telle que :

$$|\mathcal{W}_{\alpha,\beta}|_{\max} = 2t + \ell + \ell'$$

Les lettres de $\mathcal{W}_{\alpha,\beta}$ sont les $t + \ell$ premières lettres de W_α, suivies des $t + \ell'$ lettres de W_β, qui sont d'après la caractérisation des répétitions chevauchantes et maximales à gauche respectivement la répétition des d et d' premières lettres de chacune des deux répétitions. Donc le mot $\mathcal{W}_{\alpha,\beta}$ bien qu'il n'a pas de période (car les d lettres de W_α ne se répétent pas dans W_β et les d' lettres de W_β ne se répétent pas dans W_α) est la répétition des d et d' lettres a priori différentes, répétées dans chacune des deux répétitions. Le nombre de lettres a priori différentes répétées dans $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$n_{0,b_2} = d + d'$$

Plaçons nous dans le cadre général $i > 1$ (le même calcul est obtenu si $i = 1$), en tenant compte de la condition de maximalité à gauche, nous avons :

$$Y_\alpha Y_\beta = \mathbf{1}_{\left\{X_{i-1} \neq X_{i+d-1}, X_{i+\ell+t+\ell'-1} \neq X_{i+\ell+t-1}, w_i = w_{i+\ell}, w'_{i'} = w'_{i'+\ell'}\right\}}$$

où w_i et $w_{i+\ell}$ (resp. $w'_{i'}$ et $w'_{i'+\ell'}$) sont les mots qui forment le mot W_α (resp. W_β), par suite :

$$\mathbb{E}(Y_\alpha Y_\beta) = \mathbb{P}\left(X_{i-1} \neq X_{i+d-1}, X_{i+\ell+t+\ell'-1} \neq X_{i+\ell+t-1}, w_i = w_{i+\ell}, w'_{i'} = w'_{i'+\ell'}\right)$$

Soient $a_1, \ldots, a_d$ et $a'_1, \ldots, a'_d$ des éléments de $\mathcal{A}$ qui sont les valeurs prises par les lettres de $\mathcal{W}_{\alpha,\beta}$ et b un élément de $\mathcal{A}$, la maximalité à gauche des deux répétitions est donnée par :

$$b \neq a_d \qquad \text{et} \qquad \left(a_d \mathbf{1}_{\{r=0\}} + a_r \mathbf{1}_{\{r\neq 0\}}\right) \neq a'_{d'}$$

Le même calcul que celui dans la sous-section 1.3.1 (chapitre 1), donne en utilisant les probabilités conditionnelles et les propriétés des chaînes de Markov :

$$\begin{aligned}\mathbb{E}(Y_\alpha Y_\beta) &= \mathbb{P}\left(X_{i-1} \neq X_{i+d-1}, X_{i+\ell+t+\ell'-1} \neq X_{i+\ell+t-1}, w_i = w_{i+\ell}, w'_{i'} = w'_{i'+\ell'}\right) \\ &= \sum_{\substack{b,a_1,\ldots,a_d,a'_1,\ldots,a'_{d'} \in \mathcal{A} \\ b \neq a_d \\ \left(a_d \mathbf{1}_{\{r=0\}} + a_r \mathbf{1}_{\{r\neq 0\}}\right) \neq a'_{d'}}} \mu(b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\Gamma\left(\pi(a_d,a'_1)\mathbf{1}_{\{r=0\}}\right. \\ &\qquad \left. + \pi(a_r,a'_1)\mathbf{1}_{\{r\neq 0\}}\right)\left(\prod_{j=1}^{d'-1}\pi(a'_j,a'_{j+1})\right)\Gamma' \qquad (2.15)\end{aligned}$$

où Γ est donné par (1.8) et Γ' par :

$$\Gamma' = \left(\prod_{j=1}^{d'-1}\pi(a'_j,a'_{j+1})\right)^{q'-1}\left(\pi(a'_{d'},a'_1)\right)^{q'-1}\left(\pi(a'_{d'},a'_1)\prod_{j=1}^{r'-1}\pi(a'_j,a'_{j+1})\right)^{\mathbf{1}_{\{r'\neq 0\}}}$$

Posons :

$$\Phi = \mu(b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\left(\pi(a_d,a'_1)\mathbf{1}_{\{r=0\}} + \pi(a_r,a'_1)\mathbf{1}_{\{r\neq 0\}}\right)\left(\prod_{j=1}^{d'-1}\pi(a'_j,a'_{j+1})\right)$$

et :

$$\Psi = \Gamma\,\Gamma'$$

hormis les facteurs $\mu(b)$ et $\pi(a_d,a'_1)\mathbf{1}_{\{r=0\}} + \pi(a_r,a'_1)\mathbf{1}_{\{r\neq 0\}}$ de Φ, l'expression de Ψ est composée par les puissances des autres probabilités de transition qui figurent dans l'expression de Φ, l'expression (2.15) est réécrite en fonction de Φ et Ψ comme suit :

$$\mathbb{E}(Y_\alpha Y_\beta) = \sum_{\substack{b,a_1,\ldots,a_d,a'_1,\ldots,a'_{d'} \in \mathcal{A} \\ b \neq a_d \\ \left(a_d \mathbf{1}_{\{r=0\}} + a_r \mathbf{1}_{\{r\neq 0\}}\right) \neq a'_{d'}}} \Phi\,\Psi \qquad (2.16)$$

en majorant chacune des probabilités de transition dans l'expression de Ψ par ξ $(0 < \xi < 1)$, nous obtenons :

$$\begin{aligned}\Psi &\leq \xi^{(q-1)(d-1)+q-1+r\mathbf{1}_{\{r\neq 0\}}+(q'-1)(d'-1)+q'-1+r'\mathbf{1}_{\{r'\neq 0\}}} \\ &= \xi^{\left(qd-d+r\mathbf{1}_{\{r\neq 0\}}\right)+\left(q'd'-d'+r'\mathbf{1}_{\{r'\neq 0\}}\right)} \\ &= \xi^{t+(\ell-d)+t+(\ell'-d')} \\ &= \xi^{|\mathcal{W}_{\alpha,\beta}|_{\max}-n_{0,b_2}}\end{aligned}$$

comme :

$$(\ell-d)+(\ell'-d') \geq 0 \qquad \text{car} \qquad \ell-d \geq 0 \quad \text{et} \quad \ell'-d' \geq 0$$

alors :

$$|\mathcal{W}_{\alpha,\beta}|_{\max} - n_{0,b_2} \geq 2t$$

il s'en suit :

$$\xi^{|\mathcal{W}_{\alpha,\beta}|_{\max}-n_{0,b_2}} \leq \xi^{2t}$$

ainsi :

$$\Psi \leq \xi^{2t}$$

en remplaçant dans (2.16) on aura :

$$\mathbb{E}(Y_\alpha Y_\beta) \leq \left(\sum_{\substack{b,a_1,\ldots,a_d,a'_1,\ldots,a'_{d'}\in\mathcal{A} \\ b\neq a_d \\ \left(a_d\mathbf{1}_{\{r=0\}}+a_r\mathbf{1}_{\{r\neq 0\}}\right)\neq a'_{d'}}} \Phi \right) \xi^{2t} \tag{2.17}$$

et en appliquant les propriétés des chaînes de Markov pour majorer la somme de Φ dans le membre de droite de (2.17), nous avons :

• pour $r = 0$:

$$a_d\mathbf{1}_{\{r=0\}} + a_r\mathbf{1}_{\{r\neq 0\}} = a_d$$

la maximalité à gauche est donnée par $a_d \neq a'_{d'}$, l'expression entre parenthèses dans (2.17) est majorée comme suit :

$$\begin{aligned}\sum_{\substack{b,a_1,\ldots,a_d,a'_1,\ldots,a'_{d'}\in\mathcal{A} \\ b\neq a_d \\ a_d\neq a'_{d'}}} \Phi &\leq \sum_{b,a_1,\ldots,a_d,a'_1,\ldots,a'_{d'}\in\mathcal{A}} \mu(b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\pi(a_d,a'_1) \\ &\qquad \left(\prod_{j=1}^{d'-1}\pi(a'_j,a'_{j+1})\right) \\ &= 1\end{aligned}$$

• pour $r \neq 0$:

$$a_d\mathbf{1}_{\{r=0\}} + a_r\mathbf{1}_{\{r\neq 0\}} = a_r$$

la maximalité à gauche est donnée dans ce cas par $a_r \neq a'_{d'}$, de même l'expression entre parenthèses dans (2.17) est majorée comme suit :

$$\begin{aligned}\sum_{\substack{b,a_1,\ldots,a_d,a'_1,\ldots,a'_{d'}\in\mathcal{A} \\ b\neq a_d \\ a_r\neq a'_{d'}}} \Phi &\leq \sum_{b,a_1,\ldots,a_d,a'_1,\ldots,a'_{d'}\in\mathcal{A}} \mu(b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\pi(a_r,a'_1) \\ &\qquad \left(\prod_{j=1}^{d'-1}\pi(a'_j,a'_{j+1})\right) \\ &= 1\end{aligned}$$

en remplaçant dans les deux cas la majoration de la somme de Φ par 1 dans (2.17), nous obtenons :

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) \leq \xi^{2t}$$

B- Cas où la longueur $|\mathcal{W}_{\alpha,\beta}|$ est telle que $2t < |\mathcal{W}_{\alpha,\beta}| < |\mathcal{W}_{\alpha,\beta}|_{\max}$.

Contrairement au cas précédant, dans les 8 autres cas $k = 1, \ldots, 8$, au moins un mot de la répétition en β chevauche avec la répétition en α, ce qui peut être interprété en tenant compte du cas précédant, par « la répétition en β est "décalée" vers la gauche » (figures 2.5,...,2.12).

Parmi les lettres du mot $\mathcal{W}_{\alpha,\beta}$, il y a celles qui sont communes aux mots W_α et W_β, qui par ailleurs se répètent d'après les périodes des deux répétitions simultanément dans W_α et W_β, par suite dans $\mathcal{W}_{\alpha,\beta}$. Ainsi, les lettres de $\mathcal{W}_{\alpha,\beta}$ sont la répétition de n_{k,b_2} lettres a priori différentes dont au moins quelques unes sont communes aux deux répétitions.

Notons que le mot $\mathcal{W}_{\alpha,\beta}$, ne peut avoir de période, sinon d'après corollaire 2.2.1 la répétition en β ne serait pas maximale à gauche. Les n_{k,b_2} lettres a priori différentes ne peuvent donc constituer une période de $\mathcal{W}_{\alpha,\beta}$. D'après les conditions du corollaire 2.2.1 sur les d et d' lettres des répétitions en α et β, le nombre n_{k,b_2} est tel que :

$$n_{k,b_2} \leq \min(d, d') < d + d' \qquad \text{avec} \quad d \leq \ell \quad \text{et} \quad d' \leq \ell'$$

Il y a 5 cas de recouvrement à considérer :

I- Cas où W_α et W_β ont c $(1 \leq c < \min(\ell, \ell'))$ lettres communes (figure 2.15).

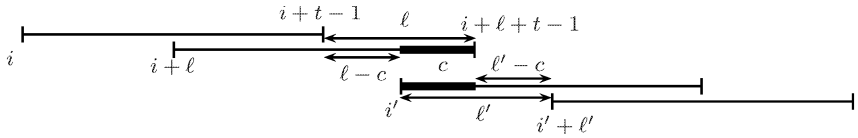

FIG. 2.15 – Le mot W_β est "déplacé" vers la gauche de c lettres $(1 \leq c < \min(\ell, \ell'))$.

Le mot W_β est "déplacé" vers la gauche de c lettres $X_{i'} \cdots X_{i'+c-1}$, qui sont à la fois les c premières et dernières lettres seulement des mots $w'_{i'}$ et $w_{i+\ell}$. La longueur du mot $\mathcal{W}_{\alpha,\beta}$ dans ce cas est telle que :

$$|\mathcal{W}_{\alpha,\beta}| = |\mathcal{W}_{\alpha,\beta}|_{\max} - c$$

En tenant compte des périodes d et d' de W_α et W_β, les c lettres contiennent les d et d' lettres, ou bien elles sont contenues parmi les d ou d' lettres.

1. **Les c lettres contiennent les d et d' lettres :** Nous avons deux sous-cas $c \geq d > d'$ et $c \geq d' > d$. Supposons par symétrie que $d' > d$ (avec d' non multiple de d), cela revient donc à étudier le cas $c \geq d' > d$ (figures 2.16, 2.17 et 2.18),

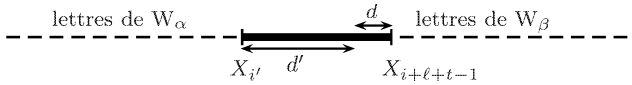

FIG. 2.16 – Cas où : $c = d + d'$.

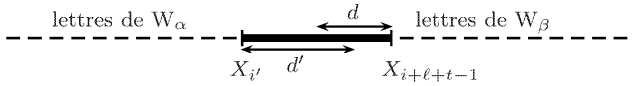

FIG. 2.17 – Cas où : $c = d + d' - z$ avec $0 < z < d$.

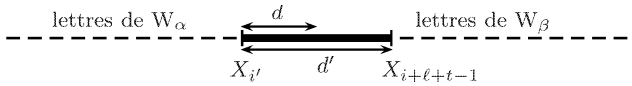

FIG. 2.18 – Cas où : $c = d'$.

Le nombre de lettres communes c est tel que :

$$c = d + d' - z \qquad \text{où} \quad 0 \leq z \leq d$$

Notons que le cas où $z = 0$ ($c = d + d'$) (figure 2.16) est exclu du fait que $X_{i'-1}$ et $X_{i'-1+d+d'}$ étant des lettres parmi les c communes, alors :

$$X_{i'-1} = X_{i'-1+d+d'} \quad \text{et} \quad X_{i'+\ell'-1} = X_{i'+d'-1} = X_{i'+d'+d-1}$$

et donc $X_{i'-1} = X_{i'+\ell'-1}$, la répétition en β n'est pas maximale à gauche.
Ce qui nous ramène donc à considérer uniquement le cas où $z \neq 0$, qui comprend deux cas, celui où $0 < z < d$, c'est à dire $c = d + d' - z$ (figure 2.17), et celui où $z = d$, c'est à dire $c = d'$ (figure 2.18).
Les g ($g \leq d$) premières lettres des c communes, sont identiques aux g dernières des d' car d est la période de W_α, comme elles sont aussi répétées parmi les dernières lettres des c car d' est la période de W_β, alors les d' lettres contiennent les d lettres de la période de W_α. Le nombre de lettres a priori différentes répétées dans $\mathcal{W}_{\alpha,\beta}$ dans ce cas est au plus égal à d, donc :

$$n_{1,b_2} \leq d$$

et on a :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{1,b_2} &\geq |\mathcal{W}_{\alpha,\beta}|_{\max} - c - d \\ &= 2t + \ell - d + \ell' - c \end{aligned}$$

ainsi :

$$|\mathcal{W}_{\alpha,\beta}| - n_{1,b_2} > 2t$$

2. **Les c lettres communes sont contenues parmi les d ou d' lettres :** Il y a 4 sous-cas $d < c \leq d'$, $d' < c \leq d$, $c \leq d < d'$ et $c \leq d' < d$ (figures 2.19, 2.20, 2.21 et 2.22),

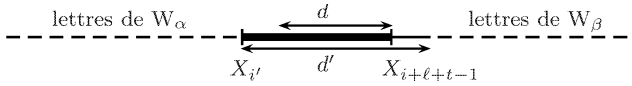

FIG. 2.19 – Cas où : $d < c \leq d'$.

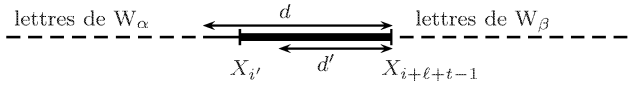

FIG. 2.20 – Cas où : $d' \leq c < d$.

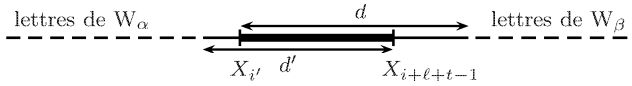

FIG. 2.21 – Cas où : $c \leq d < d'$.

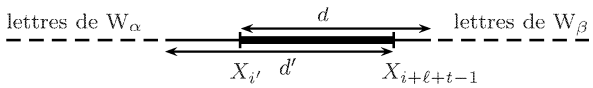

FIG. 2.22 – Cas où : $c \leq d' < d$.

(a) **Les cas $d < c \leq d'$ et $d' < c \leq d$ (figure 2.19 et 2.20) :** Ces cas étant symétriques, nous supposons alors $d < c \leq d'$. Les c lettres sont contenues parmi les d' et contiennent les d dernières de W_α (figure 2.19). Parmi les premières lettres des c lettres, et donc des d' premières de W_β, certaines se répètent avec une période égale à d car elles appartiennent à W_α. Comme $c \leq d'$, il reste $d' - c$ lettres des d' qui n'appartiennent pas à W_α mais qui sont répétées dans W_β, donc le nombre de lettres a priori différentes répétées dans $\mathcal{W}_{\alpha,\beta}$ est :

$$n_{1,b_2} = d + d' - c$$

(b) **Les cas $c \leq d < d'$ et $c \leq d' < d$ (figures 2.21 et 2.22) :** Ces cas étant symétriques, nous supposons de même que $c \leq d < d'$. Les c lettres sont contenues parmi les d dernières lettres de W_α et les d' premières lettres de la période de W_β (figure 2.21), elles sont donc les seules répétées à la fois dans W_α et W_β. Il restent $d - c$ lettres de W_α qui ne sont répétées que dans W_α, et $d' - c$ lettres de W_β qui sont répétées uniquement dans W_β, le nombre de lettres a priori différentes répétées dans $\mathcal{W}_{\alpha,\beta}$ est :

$$\begin{aligned} n_{1,b_2} &= (d - c) + c + (d' - c) \\ &= d + d' - c \end{aligned}$$

il vient de (a) et (b) que :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{1,b_2} &= |\mathcal{W}_{\alpha,\beta}|_{\max} - c - (d + d' - c) \\ &= 2t + \ell + \ell' - c - d - d' + c \\ &= 2t + (\ell - d) + (\ell' - d') \end{aligned}$$

et donc :

$$|\mathcal{W}_{\alpha,\beta}| - n_{1,b_2} > 2t$$

II- Cas où W_α et W_β ont $\ell + c$ $(1 \leq c < t - \ell)$ lettres communes (figure 2.23).

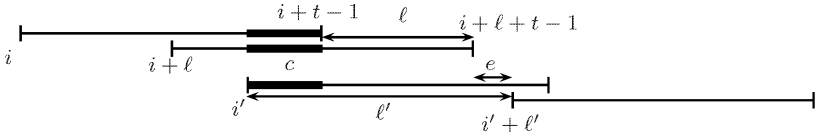

FIG. 2.23 – Le mot W_β est "déplacé" vers la gauche de $\ell + c$ lettres avec $\ell + c \leq \ell'$ et $c < \ell$.

Le mot W_β est "déplacé" vers la gauche de $\ell + c$ lettres (avec $\ell' \geq \ell + c$), où uniquement les c premières lettres de $w'_{i'}$ recouvrent les c dernières de w_i (qui appartiennent aussi à $w_{i+\ell}$). Donc le mot $w'_{i'+\ell'}$ de la répétition en β est séparé de W_α par $e = \ell' - \ell - c$ lettres.

En vertu du corollaire 2.2.1, nous avons :

$$d = \ell \qquad \text{et} \qquad d' = \ell' \quad \text{avec} \quad c < \ell$$

autrement dit, ℓ' ne peut être supérieur ou égal à un multiple de ℓ. La longueur du mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$|\mathcal{W}_{\alpha,\beta}| = |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell - c$$

Deux cas se présentent, celui où $e = 0$ (c'est dire aucune lettre ne sépare W_α du mot $w'_{i'+\ell'}$ de la répétition en β), et celui où $e \neq 0$ (il y a e lettres qui les séparent).

1. **Cas où $e = 0$ (figure 2.24) :**

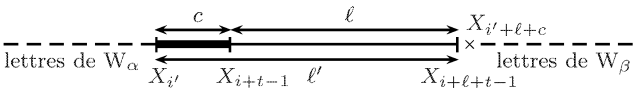

FIG. 2.24 – Cas où : $\ell' = \ell + c$ avec $c < \ell$.

Le mot $\mathcal{W}_{\alpha,\beta}$ est la répétition des ℓ lettres de la période de W_α, sans que ℓ ne soit une période de $\mathcal{W}_{\alpha,\beta}$, ainsi :

$$n_{2,b_2} = \ell$$

et nous obtenons :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{2,b_2} &= |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell - c - \ell \\ &= 2t + \ell + \ell' - \ell - c - \ell \\ &= 2t + \ell + \ell + c - \ell - c - \ell \\ &= 2t \end{aligned}$$

2. **Cas où $e \neq 0$ (figure 2.25) :**

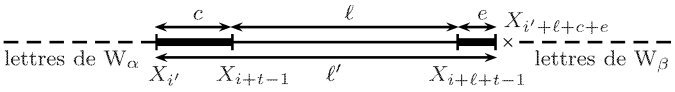

FIG. 2.25 – Cas où : $\ell' = \ell + c + e$ avec $c < \ell$.

La longueur du mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$|\mathcal{W}_{\alpha,\beta}| = |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell' + e$$

et les lettres qui le forment, sont la répétition des ℓ lettres de la période de W_α et les e lettres de W_β, ainsi :

$$n_{2,b_2} = \ell + e$$

et nous obtenons :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{2,b_2} &= |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell - c - \ell - e \\ &= 2t + \ell + \ell' - \ell - c - \ell - e \\ &= 2t + \ell + \ell + c + e - \ell - c - \ell - e \\ &= 2t \end{aligned}$$

III- Cas où W_α et W_β ont $\ell' + c$ $(1 \leq c < \ell')$ lettres communes (figure 2.26).

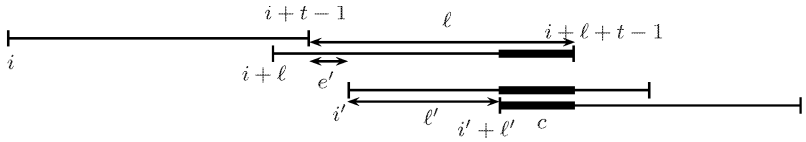

FIG. 2.26 – Le mot W_β est "déplacé" vers la gauche de $\ell' + c$ lettres avec $\ell > \ell' + c$.

Le mot W_β est "déplacé" vers la gauche de $\ell' + c$ lettres avec $\ell' + c < \ell$. Contrairement au

cas précédant les c premières lettres de $w'_{i'+\ell'}$ (qui sont aussi des lettres de $w'_{i'}$) recouvrent seulement le mot $w_{i+\ell}$ de la répétition en α, ainsi le mot w_i de la répétition en α est séparé de W_β par $e' = \ell - \ell' - c$ lettres.

De même que dans **II**, nous avons en vertu du corollaire 2.2.1 que :

$$d = \ell \qquad \text{et} \qquad d' = \ell' \quad \text{avec} \quad c < \ell'$$

c'est à dire que ℓ ne peut être un multiple de ℓ'. La longueur du mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$|\mathcal{W}_{\alpha,\beta}| = |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell' - c$$

Nous avons de même que dans **II** ci-dessus, deux cas qui se présentent $e' = 0$ (aucune lettre ne sépare le mot w_i de la répétition en α et W_β) et $e' \neq 0$ (il y a e' lettres qui les séparent).

1. **Cas où $e' = 0$ (figure 2.27) :**

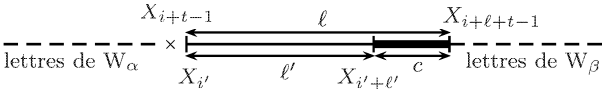

FIG. 2.27 – Cas où : $\ell = \ell' + c$ avec $c < \ell'$.

les répétitions sont formées des mêmes lettres, et on a :

$$\ell = \ell' + c \quad \text{où} \quad c < \ell'$$

La répétition des ℓ' lettres de la période de W_β forment le mot $\mathcal{W}_{\alpha,\beta}$, sans que ℓ' ne soit une période de $\mathcal{W}_{\alpha,\beta}$. Le nombre de lettres a priori différentes répétées dans $\mathcal{W}_{\alpha,\beta}$ est :

$$n_{3,b_2} = \ell'$$

le même calcul que celui dans **II-1.**, donne en tenant compte du fait que $\ell = \ell' + c$:

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{3,b_2} &= |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell' - c \\ &= 2t + \ell + \ell' - \ell' - c - \ell' \\ &= 2t + \ell' + c + \ell' - \ell' - c - \ell' \\ &= 2t \end{aligned}$$

2. **Cas où $e' \neq 0$ (figure 2.28) :**

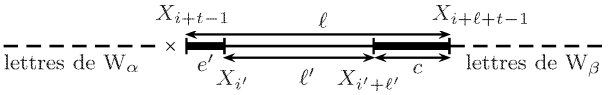

FIG. 2.28 – Cas où : $\ell = \ell' + c + e'$ avec $c < \ell'$.

les e' lettres appartiennent uniquement au mot W_α. En utilisant le même raisonnement que dans **II-2.**, le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est tel que :

$$n_{3,b_2} = \ell' + e'$$

et comme $\ell = \ell' + c + e'$, alors :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{3,b_2} &= |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell' - c - \ell' - e' \\ &= 2t + \ell + \ell' - \ell' - c - \ell' - e' \\ &= 2t + \ell' + c + e' + \ell' - \ell' - c - \ell' - e' \\ &= 2t \end{aligned}$$

IV- Cas où W_α et W_β ont $\ell + c = \ell' + c'$ ($c' < \ell$, $c < \ell'$) lettres communes (figure (2.29).

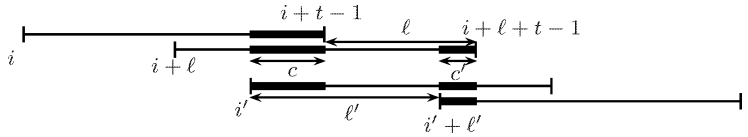

FIG. 2.29 – Le mot W_β est "décalé" vers la gauche de $\ell' + c' = \ell + c$ lettres.

Le mot W_β est "décalé" vers la gauche de $\ell' + c'$ lettres, le mot en $w'_{i'}$ de la répétition en β recouvre les mots w_i et $w_{i+\ell}$ respectivement de c et $\ell + c$ lettres. Il y a donc $\ell' - c = \ell - c'$ lettres qui séparent w_i du mot $w'_{i'+\ell'}$ de la répétition en β.

Remarquons que si ($c = \ell$ et $c' = \ell'$) ou ($c' = \ell$ et $c = \ell'$), la relation $\ell + c' = \ell' + c$ donne $\ell = \ell'$, ce qui rend d'après le corollaire 2.2.1 la répétition en β non maximale à gauche. Nous supposons donc :

$$c < \min(\ell, \ell') \qquad \text{et} \qquad c' < \min(\ell, \ell')$$

La longueur du mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$|\mathcal{W}_{\alpha,\beta}| = |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell' - c'$$

Les mots W_α et W_β sont formés de la répétition de mêmes lettres du fait de la nature du recouvrement qu'il y a entre ces mots, il en est donc de même du mot $\mathcal{W}_{\alpha,\beta}$. Nous avons $d = \ell$ et $d' = \ell'$ avec $\ell \neq \ell'$, sinon d'aprés corollaire 2.2.1, la répétition en β ne sera pas maximale à gauche.

Les lettres a priori différentes répétées dans $\mathcal{W}_{\alpha,\beta}$, sont donc à déterminer parmi les ℓ ou ℓ' lettres des périodes de W_α et W_β, leur nombre n_{4,b_2} est alors tel que :

$$n_{4,b_2} < \min(\ell, \ell') \qquad \text{avec } \ell \neq \ell'$$

Supposons par symétrie que $\ell < \ell'$, les lettres communes aux deux répétitions sont alors $X_{i'}, \ldots, X_{i'+\ell'+c'-1}$, leur nombre est tel que :

$$\ell + c = \ell' + c'$$

il est donc supérieur au nombre ℓ de lettres de la période de W_α et au nombre ℓ' de lettres de la période de W_β. Les n_{4,b_2} lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ sont à déterminer plus précisément parmi les lettres $X_{i'}, \ldots, X_{i'+\ell'+c'-1}$, car elles proviennent à la fois de W_α et W_β, de plus n_{4,b_2} n'est pas une période du mot $\mathcal{W}_{\alpha,\beta}$, car les n_{4,b_2} lettres sont répétées différemment dans W_α et W_β puisque $\ell \neq \ell'$.

De la relation $\ell + c = \ell' + c'$ et $\ell < \ell'$, nous avons $c' > c$, ainsi $c' < \ell$ (sinon on aurait $i' + \ell' \leq i + t - 1$ ce qui est vérifié uniquement pour $k = 6$, $k = 7$ et $k = 8$ (figures 2.10, 2.11 et 2.12) où $|\mathcal{W}_{\alpha,\beta}| < 2t$). Deux cas se présentent alors, $c' < c < \ell$ et $c' < \ell < c$.

1. **Cas où $c' < c < \ell$:** En tenant compte des périodes ℓ de W_α et ℓ' de W_β, les c' lettres sont répétées parmi les lettres communes (figure 2.29), elles sont telles que :

$$X_{i'} \cdots X_{i'+c'-1} = X_{i'+\ell'} \cdots X_{i'+c'-1+\ell'} = X_{i'+\ell'-\ell} \cdots X_{i'+c'-1+\ell'-\ell}$$

ainsi elles sont contenues aussi parmi les premières et les dernières lettres des c lettres, elles sont alors soit chevauchantes ou non entre elles.

(a) **Il n'y a pas de chevauchement entre les c' lettres :** Il y a au moins η ($\eta \geq 0$) lettres entre les c' premières lettres et les c' dernières, le nombre c de lettres contenant les c' lettres, est alors tel que $c = 2c' + \eta$ (figure 2.30),

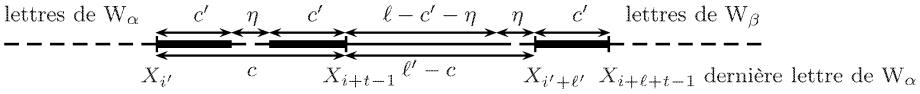

FIG. 2.30 – Cas où il n'y a pas de chevauchement entre les c' lettres.

Compte tenu de la période ℓ de la répétition en α, les c lettres sont parmi les ℓ premières de W_α du fait qu'elles sont égales aux dernières du mot W_α, qui sont aussi les premières lettres de W_β. Ces c' lettres, sont donc telles que :

$$X_{i'} \cdots X_{i'+c'-1} = X_{i'+\ell} \cdots X_{i'+c'+\ell-1}$$

pour les mêmes raisons, les η lettres sont aussi contenues parmi les c lettres, et sont telles que :

$$X_{i'+c'} \cdots X_{i'+c'+\eta'-1} = X_{i'+c'+\ell} \cdots X_{i'+c'+\eta+\ell-1}$$

car elles sont avant les c' dernières de W_α. Le nombre n_{4,b_2} de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$\begin{aligned} n_{4,b_2} &= (\ell - c' - \eta) + \eta \\ &= \ell - c' \end{aligned}$$

(b) **Il y a chevauchement entre les c' lettres :** Les c' premières lettres et les c' dernières lettres des c ont en commun γ (avec $0 < \gamma < c'$) lettres (figure 2.31),

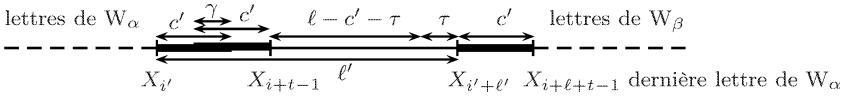

FIG. 2.31 – Cas où il y a chevauchement entre les c' lettres (on a posé $\tau = c' - \gamma$).

Comme ℓ est la période de la répétition en α, les $c' - \gamma$ premières lettres des c', sont alors telles que :

$$X_{i'} \cdots X_{i'+c'-\gamma-1} = X_{i'+\ell} \cdots X_{i'+\ell+c'-\gamma-1}$$

et donc les $c' - \gamma$ lettres figurent une seule fois parmi les $\ell - c'$ lettres, le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$\begin{aligned} n_{4,b_2} &= (\ell - c' - (c' - \gamma)) + (c' - \gamma) \\ &= \ell - c' \end{aligned}$$

2. **Cas où $c' < \ell < c$:** Il existe un entier naturel non nul y tel que $c = \ell + y$. Nous supposons que $y < \ell$ (le cas $y \geq \ell$ se déduit du cas $\ell + y$ du fait que ℓ est la période de la répétition en α). Étant donné que $\ell > c'$, entre les c' premières lettres et les c' dernières des c lettres, il y a $\nu = c - 2c'$ lettres, et en transformant cette relation après avoir remplacé c par $\ell + y$, nous obtenons $\ell - c' = \nu - (y - c')$. Comme ℓ est la période de la répétition en α, les $\nu - (y - c')$ premières lettres des ν lettres, sont telles que :

$$X_{i'+c'} \cdots X_{i'+\nu-(y-c')-1} = X_{i'+c} \cdots X_{i'+c+\ell-c'-1}$$

Remarquons là aussi que si $y = c'$ alors $c = \ell + c'$, ce qui donne en remplaçant dans l'égalité $\ell' + c' = \ell + c$ que $\ell' = 2\ell$, et dans ce cas d'après le corollaire 2.2.1, la répétition en β n'est pas maximale à gauche. Les cas qui restent à étudier sont, $y > c'$ et $y < c'$.

(a) **Si $y > c'$:** Les c' lettres sont contenues parmi les y lettres et les $y - c'$ lettres sont contenues parmi les η lettres (figure 2.32),

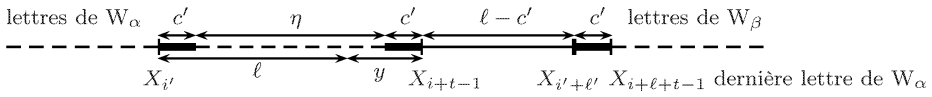

FIG. 2.32 – Cas où $c = \ell + y$ avec $0 < y < \ell$.

Là aussi deux sous-cas sont à examiner, $y - c' \geq c'$ et $y - c' < c'$.

i. Si $y - c' \geq c'$: Étant donné que ℓ est la période de W$_\alpha$, les c' premières lettres $X_{i'}, \ldots, X_{i'+c'-1}$ sont contenues parmi les $y - c'$ lettres, les $y - 2c'$ lettres qui séparent les c' premières et dernières lettres figurent une seule fois parmi les $\ell - c'$ lettres, le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est :

$$\begin{aligned} n_{4,b_2} &= (\ell - c' - (c' + (y - 2c'))) + c' + (y - c') \\ &= \ell - c' \end{aligned}$$

ii. Si $y - c' < c'$: De même que dans i., en tenant compte de la période ℓ de la répétition en α, les $y - c'$ lettres $X_{i'}, \ldots, X_{i'+y-c'-1}$ sont contenues parmi les c' premières lettres, et elles figurent une seule fois parmi les $\ell - c'$ lettres, le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$\begin{aligned} n_{4,b_2} &= (\ell - c' - (y - c')) + y - c' \\ &= \ell - c' \end{aligned}$$

(b) **Si $y < c'$ (figure 2.33) :**

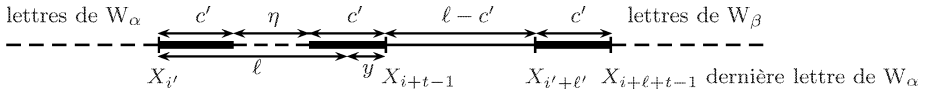

FIG. 2.33 – Cas où $c = \ell + y$, $y < c'$.

De même que dans (a), il y a deux sous-cas qui se présentent, $c' - y \geq y$ et $c' - y < y$.

i. Si $c' - y \geq y$: Comme ℓ et ℓ' sont les périodes des répétitions en α et β, les y premières lettres $X_{i'} \cdots X_{i'+y-1}$ des c' lettres sont identiques aux y dernières, qui elles, sont contenues parmi les $c' - y$ premières, par conséquent parmi les c' lettres. Donc les c' lettres sont la répétition de $y + ((c' - y) - y) = c' - y$ lettres différentes, qui ne figurent qu'une seule fois parmi les $\ell - c'$ lettres, le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est :

$$\begin{aligned} n_{4,b_2} &= (\ell - c' - (c' - y)) + (c' - y) \\ &= \ell - c' \end{aligned}$$

ii. Si $c' - y < y$: En utilisant le même argument que dans le i., comme ℓ est la période de la répétition en α, les c' lettres sont la répétition au plus de $c' - y$ lettres, et comme celles-ci ne figurent qu'une fois parmi les $\ell - c'$ lettres, le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est :

$$n_{4,b_2} = \ell - c'$$

Des cas 1. et 2., nous déduisons que le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ est tel que :

$$n_{4,b_2} = \ell - c'$$

ce qui entraine que :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{4,b_2} &= |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell' - c' - (\ell - c') \\ &= 2t + \ell + \ell' - \ell' - c' - (\ell - c') \\ &= 2t \end{aligned}$$

V- Cas où W_α et W_β ont $t+h$ $(0 < h < \min(\ell, \ell'))$ lettres communes (figure 2.34).

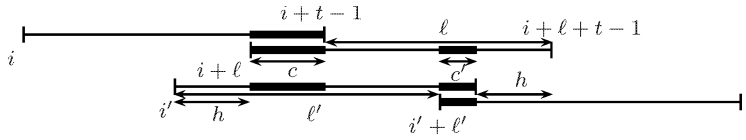

FIG. 2.34 – La répétition en β est "décalée" vers la gaughe de $t + h$ lettres.

Les mot $w'_{i'}$ et $w'_{i'+\ell'}$ de la répétition en β sont "décalés" respectivement de $\ell + c + h$ et $c' + h$ lettres où $(0 < h < \min(\ell, \ell'))$ du fait que $h = i + \ell - i'$, ainsi le mot W_β est "décalé" vers la gauche de $\ell + c + h$ lettres.

Remarquons que nous avons aussi $h = (i + \ell + t - 1) - (i' + \ell' + c' - 1)$, et donc :

$$\ell + c + h = \ell' + c' + h = t + h \qquad \text{où} \quad c = t - \ell',\ c' = t - \ell'$$

il s'en suit que :

$$\ell - c' - h = \ell' - c - h$$

et la longueur du mot $\mathcal{W}_{\alpha,\beta}$ est dans ce cas :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| &= |\mathcal{W}_{\alpha,\beta}|_{\max} - \ell - c - h \\ &= 2t + \ell + \ell' - \ell - c - h \\ &= 2t + \ell' - c - h \end{aligned}$$

De même que dans **IV**, les période d et d' des répétitions en α et β sont telles que $d = \ell$ et $d = \ell'$, et où les ℓ et ℓ' lettres ne sont pas répétées parmi les $t + h$ lettres communes aux mots W_α et W_β, sinon d'après le corollaire 2.2.1, la répétition en β ne sera pas maximale à gauche. Aussi nous écartons le cas où $\ell + \ell' = t + h$ car dans ce cas, les lettres $X_{i'-1}$ et $X_{i'+\ell'-1}$ sont identiques à la lettre $X_{i'+\ell+\ell'-1}$, qui est commune aux mots W_α et W_β, ce qui rend la répétition en β non maximale à gauche.

En écartant donc les cas que nous venons de citer, et en utilisant le même raisonnement que dans **IV**, en prenant $h+c$ et $h+c'$ au lieu de c et c', le nombre de lettres a priori différentes répétées dans le mot $\mathcal{W}_{\alpha,\beta}$ sans que ce nombre ne soit une période, est tel que :

$$n_{5,b_2} = \ell - c' - h$$

nous obtenons alors :

$$\begin{aligned} |\mathcal{W}_{\alpha,\beta}| - n_{5,b_2} &= 2t + \ell' - c - h - (\ell - c' - h) \\ &= 2t + \ell - c' - h - (\ell - c' - h) \\ &= 2t \end{aligned}$$

Remarque 2.3.1. Remarquons que dans les cas **IV** et **V**, si $\ell - c' = 1$ ou $\ell - c' - h = 1$, une seule lettre est répétée dans le mot $\mathcal{W}_{\alpha,\beta}$, et d'après le corollaire 2.2.1 la répétition en β n'est pas maximale à gauche, ce que nous excluons aussi.

De **I**, **II**, **III**, **IV** et **V**, il en résulte que pour tout $k = 1, \ldots, 5$:

$$|\mathcal{W}_{\alpha,\beta}| - n_{k,b_2} \geq 2t$$

Le nombre total de probabilités de transition d'une lettre à une autre est égal à $|\mathcal{W}_{\alpha,\beta}| - 1$, comme le mot $\mathcal{W}_{\alpha,\beta}$ est la répétition de n_{k,b_2} lettres a priori différentes, en utilisant le même principe que dans le paragraphe **A**, le produit des $|\mathcal{W}_{\alpha,\beta}| - 1$ probabilités de transition est décomposé en un produit $\Phi^*\Psi^*$ de deux quantités Φ^* et Ψ^*, où :

$$\Phi^* = \mu(b)\pi(b, a_1) \prod_{j=1}^{n_{k,b_2}-1} \pi(a_j, a_{j+1})$$

et

$$\Psi^* = \prod \underbrace{\pi(a_{n_{k,b_2}-1}, a_{n_{k,b_2}}) \cdots \pi(a_{|\mathcal{W}_{\alpha,\beta}|-n_{k,b_2}-1}, a_{|\mathcal{W}_{\alpha,\beta}|-n_{k,b_2}})}_{|\mathcal{W}_{\alpha,\beta}|-n_{k,b_2} \text{ probablités de transition}}$$

Nous avons donc, pour tout $k = 1, \ldots, 5$:

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) = \sum_{\substack{b, a_1, \ldots, a_{n_{k,b_2}} \\ \text{la répétition en } \alpha \text{ est maximale à gauche} \\ \text{la répétition en } \beta \text{ est maximale à gauche}}} \Phi^*\Psi^*$$

qui devient, en remplaçant Φ^* par son expression, et Ψ^* par le produit de toutes les majorations de chacune des probabilités de transition qui la composent par ξ $(0 < \xi < 1)$:

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) \leq \sum_{\substack{b, a_1, \ldots, a_{n_{k,b_2}} \\ \text{la répétition en } \alpha \text{ est maximale à gauche} \\ \text{la répétition en } \beta \text{ est maximale à gauche}}} \left(\mu(b)\pi(b, a_1) \prod_{j=1}^{n_{k,b_2}-1} \pi(a_j, a_{j+1}) \right) \xi^{|\mathcal{W}_{\alpha,\beta}|-n_{k,b_2}}$$

comme :

$$|\mathcal{W}_{\alpha,\beta}| - n_{k,b_2} \geq 2t \;\Rightarrow\; \xi^{|\mathcal{W}_{\alpha,\beta}|-n_{k,b_2}} \leq \xi^{2t}$$

nous obtenons pour tout $k = 1, \ldots, 5$:

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) \leq \sum_{b,a_1,\ldots,a_{n_{k,b_2}}} \left(\mu(b)\pi(b,a_j) \prod_{j=1}^{n_{k,b_2}-1} \pi(a_j,a_{j+1}) \right) \xi^{2t} \tag{2.18}$$

en utilisant les propriétés des chaînes de Markov, l'expression entre parenthèses dans (2.18) est majorée par 1, la relation (2.14) est alors vérifiée pour tout $k = 1, \ldots, 5$.

Il s'en suit de **A** et **B**, que l'expression (2.14) est satisfaite pour tout $k = 0, \ldots, 5$, ce qui donne en remplaçant dans la majoration de (2.14) dans (2.8), pour tout $k = 0, \ldots, 5$:

$$b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} \leq |\mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}}|\xi^{2t} \tag{2.19}$$

où $|\mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}}|$ désigne le nombre de positions (α,β) des mots W_α et W_β pour les lesquelles $|\mathcal{W}_{\alpha,\beta}| > 2t$.

Il reste donc d'après (2.19) à majorer $|\mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}}|$. Nous avons pour tout $k = 0, \ldots, 5$ et pour $\alpha = (i, i+\ell)$ fixé :

- la position i', a au plus $2t$ positions et ℓ' a $t-1$ valeurs possibles, ainsi β a au plus $2t(t-1)$ positions,
- la position i, a au plus n positions du fait de la longueur de la séquence S qui est n, et ℓ ayant $t-1$ valeurs possibles,

nous avons donc au plus $2nt(t-1)^2$ positions pour α et β, il s'en suit alors en majorant $t-1$ par t, que pour tout $k = 0, \ldots, 5$:

$$|\mathrm{G}_{k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}}| \leq 2nt^3$$

ce qui donne en remplaçant dans (2.19) :

$$\begin{aligned} b_{2,k,\{|\mathcal{W}_{\alpha,\beta}|>2t\}} &\leq 2nt^3\xi^{2t} \\ &= 2\left(\frac{t}{n}\right)^3 \left(n^2\xi^t\right)^2 \end{aligned} \tag{2.20}$$

et en faisant la somme de $k = 0$ à $k = 5$ dans (2.20), nous obtenons (2.13). C.Q.F.D

2$^{\mathrm{e}}$cas : la longueur $|\mathcal{W}_{\alpha,\beta}|$ du mot $\mathcal{W}_{\alpha,\beta}$ est telle que $|\mathcal{W}_{\alpha,\beta}| < 2t$.

Les mots $w'_{i'}$ et $w'_{i'+\ell'}$ de la répétition en β recouvrent le mot w_i, et donc $w_{i+\ell}$ de la répétition en α (figures 2.10, 2.11 et 2.12). Le nombre de lettres communes c aux mots W_α et W_β est compris entre 3 et $2t-2$ qui sont respectivement les valeurs minimale et maximale, obtenues pour $\ell + \ell' + 1$ avec $\ell = \ell' = 1$, et $t + \min(\ell, \ell')$ avec $\min(\ell, \ell') = t - 2$.

Pour $i < i'$ nous avons $i' + \ell' \leq i + t - 1$, les d ($d \leq \ell$) et d' ($d' \leq \ell'$) lettres sont donc répétées parmi les c lettres communes, ce qui rend d'après le corollaire 2.2.2, la répétition en β non maximale à gauche. Il reste donc à étudier le cas où $i = i'$ qui est réalisé pour $k = 7$ ou $k = 8$ (figures 2.11 et 2.12). La longueur du mot $\mathcal{W}_{\alpha,\beta}$ est telle que :

$$|\mathcal{W}_{\alpha,\beta}| = t + \max(\ell, \ell') \qquad \text{où} \quad \max(\ell, \ell') < t \quad \text{et} \quad \ell \neq \ell'$$

Supposons par symétrie que $\ell' > \ell$ et donc $i + \ell < i' + \ell'$ (figure 2.35), la longueur du mot $\mathcal{W}_{\alpha,\beta}$ dans ce cas est précisément :

$$|\mathcal{W}_{\alpha,\beta}| = t + \ell'$$

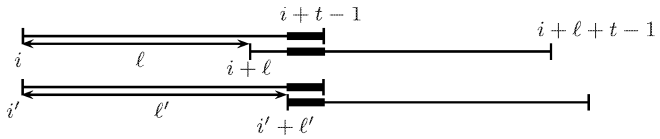

FIG. 2.35 – Cas où $i = i' < i + \ell < i' + \ell' \leq i + t - 1$.

Le mot $\mathcal{W}_{\alpha,\beta}$ est alors formé des deux répétitions en $\alpha = (i, i + \ell)$ et $\beta = (i, i + \ell')$, et en utilisant le fait que ℓ et ℓ' sont les périodes de W_α et W_β avec $\ell < \ell'$, les ℓ premières lettres de $\mathcal{W}_{\alpha,\beta}$ vérifient la relation suivante :

$$X_i = X_{i+\ell} = X_{i+\ell'} = X_{i+(\ell'-\ell)}$$

qui en plus sont telles que :

- si $\ell' - \ell < \ell$, les lettres sont égales à $X_{i+(\ell'-2\ell)}$,
- si $\ell' - \ell < \ell$, les lettres sont égales à $X_{i+(2\ell-\ell')}$,

En continuant le même raisonnement, cela revient à déterminer le p.g.c.d $\ell \wedge \ell'$ de ℓ et ℓ' en utilisant l'algorithme d'Euclide. Soit $s = \ell \wedge \ell'$, nous avons donc :

$$\begin{cases} X_i = X_{i+s} = \cdots = X_{i+\ell} = X_{i+\ell+s} = \cdots = X_{i+\ell'} \\ \quad \vdots \\ X_{i+s-1} = X_{i+2s-1} = \cdots = X_{i+\ell+s-1} = X_{i+\ell+2s-1} = \cdots = X_{i+\ell'+s-1} \end{cases}$$

Ainsi les s premières lettres a priori différentes du mot w_i qui est commun à W_α et W_β, forment la période du mot $\mathcal{W}_{\alpha,\beta}$. Les premières composantes des positions α et β de W_α et W_β étant les mêmes, nous supposons donc par la suite que la répétition en α (et donc aussi la répétition en β) est maximale à gauche.

Soit $\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$ la partie de $\mathcal{I}$ telle que β appartient au voisinage B_α de α avec $\beta \neq \alpha$, elle est définie par :

$$\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \left\{\alpha, \beta \in \mathcal{I} \,/\, i = i',\, \ell < \ell',\, \ell \wedge \ell' = s \geq 1\right\}$$

Nous avons :

$$b_{2,7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \sum_{\alpha,\beta \in \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}} \mathbb{E}\left(Y_\alpha Y_\beta\right) \tag{2.21}$$

et nous avons aussi par symétrie :

$$\mathrm{G}_{8,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \left\{\alpha, \beta \in \mathcal{I} \,/\, i = i',\, \ell > \ell',\, \ell \wedge \ell' = s \geq 1\right\}$$

et

$$b_{2,8,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \sum_{\alpha,\beta \in \mathrm{G}_{8,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}} \mathbb{E}\left(Y_\alpha Y_\beta\right)$$

en excluant le cas $k = 6$ où la répétition en β n'est pas maximale à gauche, alors :

$$b_{2,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = b_{2,7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} + b_{2,8,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$$

Lemme 2.3.2. *Si la longueur du mot $\mathcal{W}_{\alpha,\beta}$ est telle que $|\mathcal{W}_{\alpha,\beta}| < 2t$. Si $t = o(n)$, il existe une constante positive $C(\xi)$ telle que :*

$$b_{2,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} \leq \left(\frac{2\,C(\xi)}{n}\right) n^2 \xi^t \tag{2.22}$$

Démonstration. Le même raisonnement utilisé pour majorer $b_{2,7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$ donne par symétrie la même majoration pour $b_{2,8,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$, il suffit donc de majorer $b_{2,7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$ dont l'expression est donnée dans (2.21).

Soient $a_1, \ldots, a_s$ des éléments de $\mathcal{A}$, qui sont les valeurs prises par les s lettres répétées dans le mot $\mathcal{W}_{\alpha,\beta}$, nous avons d'après la caractérisation des répétitions chevauchantes et maximales à gauche, en posant $t + \ell' = qs + r$ avec $0 \leq r < s$:

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) = \sum_{\substack{b,a_1,\ldots,a_s \in \mathcal{A} \\ b \neq a_s}} \mu(b)\pi(b,a_j) \left(\prod_{j=1}^{s-1} \pi(a_j, a_{j+1})\right) \Gamma'' \tag{2.23}$$

où :

$$\Gamma'' = \left(\prod_{j=1}^{s-1} \pi(a_j, a_{j+1})\right)^{q-1} (\pi(a_s, a_1))^{q-1} \left(\pi(a_s, a_1) \prod_{j=1}^{r-1} \pi(a_j, a_{j+1})\right)^{\mathbf{1}_{\{r\neq 0\}}}$$

En utilisant le même procédé de calcul pour majorer l'expression (1.8) dans la démonstration du lemme 1.3.2 (chapitre 1), nous avons en majorant chacune des probabilités de transition dans Γ'' par ξ $(0 < \xi < 1)$, tout en tenant compte de la division euclidienne de $t + \ell'$ par s donnée ci-dessus :

$$\Gamma'' \leq \xi^{t+\ell'-s}$$

en remplaçant cette dernière dans (2.23), nous obtenons :

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) \leq \left(\sum_{\substack{b,a_1,\ldots,a_s \in \mathcal{A} \\ b \neq a_s}} \mu(b)\pi(b,a_j) \left(\prod_{j=1}^{s-1} \pi(a_j, a_{j+1})\right)\right) \xi^{t+\ell'-s}$$

et comme d'après les propriétés des chaînes de Markov :

$$\sum_{\substack{b,a_1,\ldots,a_s \in \mathcal{A} \\ b \neq a_s}} \mu(b)\pi(b,a_j) \left(\prod_{j=1}^{s-1} \pi(a_j, a_{j+1})\right) \leq 1$$

alors :

$$\mathbb{E}\left(Y_\alpha Y_\beta\right) \leq \xi^{t+\ell'-s} \tag{2.24}$$

La majoration dans (2.24) dépend de s (donc de ℓ et ℓ' car $s = \ell \wedge \ell'$), il en sera alors de même d'après (2.21) de $b_{2,7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$.

Comme $s > 1$, nous avons deux cas :

- celui où ℓ et ℓ' sont premiers entre eux, ce qui correspond à $s = 1$;
- celui où ℓ et ℓ' ne sont pas premiers entre eux, ce qui correspond à $s \geq 2$.

Ce qui nous ramène à partitionner l'ensemble $\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$ en deux sous-ensembles comme suit :

$$\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} = \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s=1\}} \cup \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s\geq 2\}}$$

où :

$$\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s=1\}} = \{\alpha,\beta \in \mathcal{I} \,/\, i=i',\, \ell<\ell',\, \ell\wedge\ell'=s=1\}$$

et

$$\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s\geq 2\}} = \{\alpha,\beta \in \mathcal{I} \,/\, i=i',\, \ell<\ell',\, \ell\wedge\ell'=s\geq 2\}$$

ainsi l'expression (2.21) réécrite en y remplaçant (2.24), devient :

$$b_{2,7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}} \leq \sum_{\alpha,\beta\in \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s=1\}}} \xi^{t+\ell'-s} + \sum_{\alpha,\beta\in \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s\geq 2\}}} \xi^{t+\ell'-s} \qquad (2.25)$$

et la majoration de $b_{2,7,\{|\mathcal{W}_{\alpha,\beta}|<2t\}}$ est déduite alors des majorations des quantités :

$$\sum_{\alpha,\beta\in \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s=1\}}} \xi^{t+\ell'-s} \quad \text{et} \quad \sum_{\alpha,\beta\in \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s\geq 2\}}} \xi^{t+\ell'-s}$$

qui représentent respectivement le premier terme et le deuxième terme du membre de droite de l'expresion (2.25).

• **Majoration du premier terme.**

Dans ce terme la somme se fait pour tout ℓ et ℓ' premiers entre eux ($s=1$), une seule lettre est donc répétée dans le mot $\mathcal{W}_{\alpha,\beta}$. Donc i a $n-t-2$ positions possibles, par suite on a :

$$\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s=1\}} = \bigcup_{i=1}^{n-t-2} \bigcup_{\substack{\ell<\ell' \\ \ell\wedge\ell'=1}} \{\alpha,\beta\in\mathcal{I}\}$$

comme $\ell,\ell' = 1,\ldots,t-1$, dans la somme sur $\mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s=1\}}$, les couples (α,β) de $\mathcal{I}^2$ sont tels que $\ell\wedge\ell'=1$ avec $\ell,\ell'=1,\ldots,t-1$, la somme du premier terme est donnée par :

$$\begin{aligned}
\sum_{\alpha,\beta\in \mathrm{G}_{7,\{|\mathcal{W}_{\alpha,\beta}|<2t\},\{s=1\}}} \xi^{t+\ell'-1} &= \sum_{i=1}^{n-t-2} \sum_{\substack{\ell<\ell' \\ \ell\wedge\ell'=1}} \xi^{t+\ell'-1} \\
&= \sum_{i=1}^{n-t-2} \sum_{\substack{\ell,\ell'=1,\ldots,t-1 \\ \ell<\ell' \\ \ell\wedge\ell'=1}} \xi^{t+\ell'-1} \\
&= (n-t-2)\xi^t \left(\sum_{\substack{\ell=1,\ldots,t-2 \\ \ell'=\ell+1,\ldots,t-1 \\ \ell\wedge\ell'=1}} \xi^{\ell'-1} \right) \\
&= n\,\xi^t \left(1-\frac{t}{n}-\frac{2}{n}\right) \left(\sum_{\substack{\ell=1,\ldots,t-2 \\ \ell'=\ell+1,\ldots,t-1 \\ \ell\wedge\ell'=1}} \xi^{\ell'-1} \right)
\end{aligned}$$

Pour $t = o(n)$, nous avons :

$$\sum_{\alpha,\beta\in \mathrm{G}_{7,\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\},\{s=1\}}} \xi^{t+\ell'-1} \leq \left(\sum_{\substack{\ell=1,\ldots,t-2 \\ \ell'=\ell+1,\ldots,t-1 \\ \ell\wedge\ell'=1}} \xi^{\ell'-1} \right) n\,\xi^t \tag{2.26}$$

cela revient donc à majorer le facteur entre parenthèses du membre de droite (2.26) pour tout ℓ, $\ell' = 1, \ldots, t-1$ avec $\ell < \ell'$.

Nous avons :

$$\begin{aligned} \sum_{\substack{\ell=1,\ldots,t-2 \\ \ell'=\ell+1,\ldots,t-1 \\ \ell\wedge\ell'=1}} \xi^{\ell'-1} &\leq \sum_{\ell=1}^{t-2} \sum_{\ell'=\ell+1}^{t-1} \xi^{\ell'-1} \\ &= \frac{1}{1-\xi} \left(\sum_{\ell=1}^{t-2} \xi^{\ell} - \sum_{\ell=1}^{t-2} \xi^{t-1} \right) \\ &= \left(\frac{\xi}{(1-\xi)^2} \right) \left(1 + (t-2)\xi^{t-1} - (t-1)\xi^{t-2} \right) \end{aligned}$$

comme $0 < \xi < 1$, alors pour tout $t \geq 2$:

$$(t-2)\xi^{t-1} - (t-1)\xi^{t-2} \leq 0$$

par suite :

$$1 + (t-2)\xi^{t-1} - (t-1)\xi^{t-2} \leq 1$$

ainsi :

$$\sum_{\substack{\ell=1,\ldots,t-2 \\ \ell'=\ell+1,\ldots,t-1 \\ \ell\wedge\ell'=1}} \xi^{\ell'-1} \leq \frac{\xi}{(1-\xi)^2}$$

en posant :

$$C_1(\xi) = \frac{\xi}{(1-\xi)^2}$$

que l'on remplaçera dans (2.26), le premier terme est donc majoré comme suit :

$$\sum_{\alpha,\beta\in \mathrm{G}_{7,\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\},\{s=1\}}} \xi^{t+\ell'-1} \leq C_1(\xi)\, n\,\xi^t \tag{2.27}$$

• Majoration du deuxième terme.

Les lettres du mot $\mathcal{W}_{\alpha,\beta}$ sont la répétition des s ($s \geq 2$) premières lettres. Dans le deuxième terme les couples (α, β) sont tels que ℓ et ℓ' ne sont pas premiers entre eux ($s \geq 2$). La plus petite valeur de s dans ce cas, qui est aussi celle de ℓ, est 2 (car ℓ a pour plus petite valeur s). Comme $\ell \wedge \ell' = 2$, la plus petite valeur de ℓ' est $\ell + s = 4$, dans ce cas i a $n - t - 4$ positions possibles, ainsi :

$$\mathrm{G}_{7,\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\},\{s\geq 2\}} = \bigcup_{i=1}^{n-t-4} \bigcup_{\substack{\ell<\ell' \\ \ell\wedge\ell'=s\geq 2}} \{\alpha, \beta \in \mathcal{I}\}$$

Le deuxième terme de (2.25) s'écrit donc :

$$\sum_{\alpha,\beta\in \mathrm{G}_7,\left\{\left|\mathcal{W}_{\alpha,\beta}\right|<2t\right\},\{s\geq 2\}} \xi^{t+\ell'-s} = \sum_{i=1}^{n-t-4} \sum_{\substack{\ell,\ell'=1,\dots,t-1\\ \ell<\ell'\\ \ell\wedge\ell'=s\geq 2}} \xi^{t+\ell'-s}$$

$$= (n-t-4)\,\xi^t \left(\sum_{\substack{\ell,\ell'=1,\dots,t-1\\ \ell<\ell'\\ \ell\wedge\ell'=s\geq 2}} \xi^{\ell'-s} \right)$$

$$= \left(1-\frac{t}{n}-\frac{4}{n}\right) n\,\xi^t \left(\sum_{\substack{\ell,\ell'=1,\dots,t-1\\ \ell<\ell'\\ \ell\wedge\ell'=s\geq 2}} \xi^{\ell'-s} \right)$$

Pour $t = o(n)$, nous avons :

$$\sum_{\alpha,\beta\in \mathrm{G}_7,\left\{\left|\mathcal{W}_{\alpha,\beta}\right|<2t\right\},\{s\geq 2\}} \xi^{t+\ell'-s} \leq \left(\sum_{\substack{\ell,\ell'=1,\dots,t-1\\ \ell<\ell'\\ \ell\wedge\ell'=s\geq 2}} \xi^{\ell'-s} \right) n\,\xi^t \tag{2.28}$$

ce qui nous ramène donc aussi à majorer le facteur entre parenthèses du membre de droite de l'expression (2.28).

La plus grande valeur de s étant ℓ, alors de $\ell\wedge\ell' = \ell$ pour ℓ ($\ell \leq t-1$) assez grand nous déduisons que $\ell' = 2\ell = 2s$, comme $\ell' \leq t-1$, alors $s \leq \left[\frac{t-1}{2}\right]$ (où $[\,\cdot\,]$ désigne la partie entière). Si nous notons par N et u ($0 \leq u < s$) respectivement le quotient et le reste de la division euclidienne de ℓ' pas s, alors ℓ' prend pour valeurs tous les multiples de s, de $\ell+s$ à $t-u = Ns$. Nous avons ainsi :

$$\sum_{\substack{\ell,\ell'=1,\dots,t-1\\ \ell<\ell'\\ \ell\wedge\ell'=s\geq 2}} \xi^{\ell'-s} = \sum_{s=2}^{\left[\frac{t-1}{2}\right]} \sum_{\frac{\ell}{s}=1}^{N-1} \sum_{\frac{\ell'}{s}=\frac{\ell}{s}+1}^{N} (\xi^s)^{\frac{\ell'}{s}-1}$$

$$= \sum_{s=2}^{\left[\frac{t-1}{2}\right]} \sum_{\frac{\ell}{s}=1}^{N-1} \frac{\xi^{\ell}-\xi^{Ns}}{1-\xi^s}$$

$$= \sum_{s=2}^{\left[\frac{t-1}{2}\right]} \frac{\xi^s}{(1-\xi^s)^2} \left(1+(N-1)\xi^{Ns}-N\,\xi^{(N-1)s}\right)$$

comme $0<\xi<1$, alors pour tout $N \geq 2$:

$$(N-1)\,\xi^{Ns} \leq N\,\xi^{(N-1)s}$$

par suite :

$$1+(N-1)\xi^{Ns}-N\,\xi^{(N-1)s}\leq 1$$

ainsi :

$$\begin{aligned}\sum_{\substack{\ell,\ell'=1,\ldots,t-1\\ \ell<\ell'\\ \ell\wedge\ell'=s\geq 2}}\xi^{\ell'-s} &\leq \sum_{s=2}^{\left[\frac{t-1}{2}\right]}\frac{\xi^s}{(1-\xi^s)^2}\\ &\leq \sum_{s=2}^{+\infty}\frac{\xi^s}{(1-\xi^s)^2}\end{aligned}$$

comme cette série est d'après le critère de d'Alembert convergente, en posant :

$$C_2(\xi)=\sum_{s=2}^{+\infty}\frac{\xi^s}{(1-\xi^s)^2}$$

nous obtenons en remplaçant $C_2(\xi)$ dans (2.28) :

$$\sum_{\alpha,\beta\in \mathrm{G}_{7,\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\},\{s\geq 2\}}}\xi^{t+\ell'-s}\leq C_2(\xi)\,n\,\xi^t \tag{2.29}$$

En remplaçant de nouveau (2.27) et (2.29) dans (2.25), nous obtenons :

$$\begin{aligned}b_{2,7,\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\}} &\leq C_1(\xi)\,n\,\xi^t+C_2(\xi)\,n\,\xi^t\\ &=\left(\frac{C_1(\xi)+C_2(\xi)}{n}\right)n^2\,\xi^t\end{aligned}$$

et en posant :

$$C(\xi)=C_1(\xi)+C_2(\xi)$$

nous obtenons :

$$b_{2,7,\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\}}\leq\left(\frac{C(\xi)}{n}\right)n^2\xi^t \tag{2.30}$$

étant donné que par symétrie, nous obtenons aussi :

$$b_{2,8,\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\}}\leq\left(\frac{C(\xi)}{n}\right)n^2\xi^t \tag{2.31}$$

la majoration dans (2.22) est alors déduite en additionnant membres à membres les expressions de (2.30) et (2.31). C.Q.F.D

Proposition 2.3.2. *Pour $t=o(n)$, il existe une constante positive $C(\xi)$ telle que :*

$$b_2\leq 24\left(\frac{t}{n}\right)^3\left(n^2\xi^t\right)^2+\left(\frac{4C(\xi)}{n}\right)n^2\xi^t$$

Démonstration. En appliquant le lemme 2.3.1 et lemme 2.3.2, la majoration de b_2 est obtenue en remplaçant dans (2.12), les quantités $b_{2,\left\{|\mathcal{W}_{\alpha,\beta}|>2t\right\}}$ et $b_{2\left\{|\mathcal{W}_{\alpha,\beta}|<2t\right\}}$ par leurs majorations dans (2.13) et (2.22). C.Q.F.D

2.3.3 Majoration de b_3

Nous commençons par rappeler l'expression de b_3 :

$$b_3 = \sum_{\alpha \in \mathcal{I}} \mathbb{E}\left|\left[\mathbb{E}(Y_\alpha - \mathbb{E}(Y_\alpha)|\sigma\left(Y_\beta\,;\ \beta \notin \mathrm{B}_\alpha\right)\right]\right|$$

D'après l'expression de b_3, les positions α et β des répétitions en α et β sont non voisines (c'est à dire qu'il y a au moins une lettre qui sépare les deux répétitions). Étant donné qu'il s'agit de majorer b_3, plaçons nous dans le cas où une seule lettre sépare la répétition en α de celle en β (figure 2.36), car ce cas contient tous les autres.

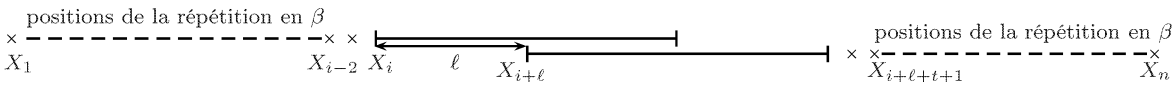

FIG. 2.36 – Positions des répétitions en β non voisines de celle en α.

Supposons que la répétition en α occure à la position $\alpha = (i, i+\ell)$, la position β étant non voisine de α, alors la répétition en β est générée par les variables aléatoires autres que celles de la répétition en α. Ces variables aléatoires sont $X_1, \ldots, X_{i-2}, X_{i+t+\ell+1}, \ldots, X_n$ (figure 2.36), la σ-algebre $\sigma\left(Y_\beta\,;\ \beta \notin \mathrm{B}_\alpha\right)$ est donc telle que :

$$\sigma\left(Y_\beta\,;\ \beta \notin \mathrm{B}_\alpha\right) \subset \sigma\left(X_1, \ldots, X_{i-2}, X_{i+\ell+t+1}, \ldots, X_n\right)$$

Proposition 2.3.3. *Pour $t = o(n)$ nous avons :*

$$b_3 \leq \left(\frac{t}{n^3}\right)\left(n^2\xi^t\right)^2 + \left(\frac{t}{n}\right)n^2\xi^t$$

Démonstration. En vertu des propriétés de l'espérance conditionnelle par rapport à une σ-algèbre (appliquées pour $\sigma\left(Y_\beta\,;\ \beta \notin \mathrm{B}_\alpha\right)$ et $\sigma\left(X_1, \ldots, X_{i-2}, X_{i+\ell+t+1}, \ldots, X_n\right)$ vérifiant la relation d'inclusion ci-dessus) et des chaînes de Markov, nous avons :

$$\begin{aligned}
&\left|\mathbb{E}[\mathbb{E}(Y_\alpha - \mathbb{E}(Y_\alpha)|\sigma\left(Y_\beta\,;\ \beta \notin \mathrm{B}_\alpha\right)]\right| \\
&\quad = \left|\mathbb{E}[\mathbb{E}(Y_\alpha - \mathbb{E}\left(\mathbb{E}(Y_\alpha)|\sigma\left(X_1, \ldots, X_{i-2}, X_{i+\ell+t+1}, \ldots, X_n\right)\right)|\sigma\left(Y_\beta\,;\ \beta \notin \mathrm{B}_\alpha\right)]\right| \\
&\quad = \left|\mathbb{E}[\mathbb{E}\left(Y_\alpha - \mathbb{E}\left(Y_\alpha|X_{i-2}, X_{i+\ell+t+1}\right)\right)]\right|
\end{aligned}$$

il s'en suit :

$$\begin{aligned}
\left|\mathbb{E}[\mathbb{E}\left(Y_\alpha - \mathbb{E}\left(Y_\alpha|X_{i-2}, X_{i+\ell+t+1}\right)\right)]\right| &\leq \mathbb{E}\left|\left[\mathbb{E}\left(Y_\alpha\right) - \mathbb{E}\left(Y_\alpha|X_{i-2}, X_{i+\ell+t+1}\right)\right]\right| \\
&= \mathbb{E}\left|\left[\mathbb{E}\left(Y_\alpha|X_{i-2}, X_{i+\ell+t+1}\right) - \mathbb{E}\left(Y_\alpha\right)\right]\right|
\end{aligned}$$

et en remplaçant dans l'expression de b_3, nous obtenons :

$$b_3 \leq \sum_{\alpha \in \mathcal{I}} \mathbb{E}\left|\left[\mathbb{E}\left(Y_\alpha|X_{i-2}, X_{i+\ell+t+1}\right) - \mathbb{E}\left(Y_\alpha\right)\right]\right| \tag{2.32}$$

ce qui nous ramène ainsi à commencer par majorer en premier, pour tout α dans $\mathcal{I}$, le terme $\mathbb{E}\left|\left[\mathbb{E}\left(Y_\alpha|X_{i-2}, X_{i+\ell+t+1}\right) - \mathbb{E}\left(Y_\alpha\right)\right]\right|$ de la somme du membre de droite l'inégalité (2.32).

Nous avons d'après la définition de l'espérance mathématique :

$$\begin{aligned}
&\mathbb{E}\left|\mathbb{E}[Y_\alpha|X_{i-2}, X_{i+\ell+t+1}] - \mathbb{E}(Y_\alpha)\right| \\
&\quad = \sum_{x,y\in\mathcal{A}} \left|[\mathbb{E}(Y_\alpha|X_{i-2}=x, X_{i+\ell+t+1}=y) - \mathbb{E}(Y_\alpha)]\right| \mathbb{P}(X_{i-2}=x, X_{i+\ell+t+1}=y) \\
&\quad = \sum_{x,y\in\mathcal{A}} |\mathbb{E}(Y_\alpha|X_{i-2}=x, X_{i+\ell+t+1}=y)\mathbb{P}(X_{i-2}=x, X_{i+\ell+t+1}=y) \\
&\qquad -\mathbb{E}(Y_\alpha)\mathbb{P}(X_{i-2}=x, X_{i+\ell+t+1}=y)| \qquad (2.33)
\end{aligned}$$

en posant :

$$P_1 = \mathbb{E}(Y_\alpha|X_{i-2}=x, X_{i+\ell+t+1}=y)\mathbb{P}(X_{i-2}=x, X_{i+\ell+t+1}=y)$$

et

$$P_2 = \mathbb{E}(Y_\alpha)\mathbb{P}(X_{i-2}=x, X_{i+\ell+t+1}=y)$$

il en résulte alors, en remplaçant dans (2.33) :

$$\begin{aligned}
|\mathbb{E}[Y_\alpha|X_{i-2}, X_{i+\ell+t+1}] - \mathbb{E}(Y_\alpha)| &= \sum_{x,y\in\mathcal{A}} |P_1 - P_2| \\
&\leq \sum_{x,y\in\mathcal{A}} P_1 + P_2 \qquad (2.34)
\end{aligned}$$

Ainsi pour majorer $\mathbb{E}\left|[\mathbb{E}\left(Y_\alpha|X_{i-2}, X_{i+\ell+t+1}\right) - \mathbb{E}\left(Y_\alpha\right)]\right|$, il suffit donc d'après (2.34) de majorer les expressions de P_1 et P_2.

• **Majoration de P_1.**

Pour tout $b, a_1, \ldots, a_d, c$ dans $\mathcal{A}$ tels que $b \neq a_1$ et pour tout x et y dans $\mathcal{A}$, nous avons :

$$\begin{aligned}
P_1 = \sum_{\substack{b,a_1,\ldots,a_d,c\in\mathcal{A} \\ b\neq a_d}} &\mathbb{P}(X_{i-2}=x, X_{i-1}=b, X_i=a_1, \ldots, X_{i+\ell+t-1}=a_d\,\mathbf{1}_{\{r=0\}} + a_r\,\mathbf{1}_{\{r\neq 0\}}, \\
&X_{i+\ell+t}=c, X_{i+\ell+t+1}=y)
\end{aligned}$$

D'après la définition des chaînes de Markov, et la caractérisation des répétitions chevauchantes et maximales à gauche appliquée pour le mot W_α ayant pour période d et longueur $t+\ell$ telle que $t+\ell = qd+r$ $(0 \leq r < d)$, l'expression de P_1 est réécrite comme suit :

$$\begin{aligned}
P_1 = \sum_{\substack{b,a_1,\ldots,a_d,c\in\mathcal{A} \\ b\neq a_d}} &\mu(x)\pi(x,b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\Gamma\,\pi\left(a_r\,\mathbf{1}_{\{r\neq 0\}} + a_d\,\mathbf{1}_{\{r=0\}}, c\right) \\
&\times \pi(c,y) \qquad (2.35)
\end{aligned}$$

où Γ est donnée par (1.8).

En majorant chacune des probabilités de transition par ξ $(0 < \xi < 1)$, nous avons $\Gamma \leq \xi^t$ (démonstration du lemme 1.3.2 (chapitre 1)) que nous remplaçons dans (2.35), qui devient alors :

$$P_1 \leq \Delta_{x,y}\,\xi^t \qquad (2.36)$$

où $\Delta_{x,y}$ est donnée par :

$$\Delta_{x,y} = \sum_{\substack{b,a_1,\ldots,a_d,c\in\mathcal{A}\\ b\neq a_d}} \mu(x)\pi(x,b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\pi\left(a_r\,\mathbf{1}_{\{r\neq 0\}}+a_d\,\mathbf{1}_{\{r=0\}},c\right) \times\pi(c,y) \tag{2.37}$$

• **Majoration de P_2.**

De même que ci-dessus pour tout $b,a_1,\ldots,a_d,c$ dans $\mathcal{A}$ tels que $b\neq a_1$ et pour tout x et y dans $\mathcal{A}$, nous avons :

$$P_2=\mathbb{E}(Y_\alpha)\sum_{\substack{b,a_1,\ldots,a_d,c\in\mathcal{A}\\ b\neq a_d}}\mathbb{P}(X_{i-2}=x,X_{i-1}=b,X_i=a_1,\ldots,X_{i+\ell+t-1}=a_d\,\mathbf{1}_{\{r=0\}}+a_r\,\mathbf{1}_{\{r\neq 0\}},$$
$$X_{i+\ell+t}=c,X_{i+\ell+t+1}=y)$$

en utilisant les mêmes arguments que ci-dessus, P_2 s'écrira :

$$P_2=\mathbb{E}(Y_\alpha)\sum_{\substack{b,a_1,\ldots,a_d,c\in\mathcal{A}\\ b\neq a_d}}\mu(x)\pi(x,b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\Gamma\,\pi\left(a_r\,\mathbf{1}_{\{r\neq 0\}}+a_d\,\mathbf{1}_{\{r=0\}},c\right) \times\pi(c,y) \tag{2.38}$$

où Γ est donnée par (1.8), en appliquant (1.11) du lemme 1.3.2 (chapitre 1) et (2.37), à (2.38), nous obtenons :

$$P_2\leq\Delta_{x,y}\,\xi^{2t} \tag{2.39}$$

où $\Delta_{x,y}$ est donnée par (2.37)

En remplaçant (2.38) et (2.36) dans (2.34), celle-ci devient :

$$|\mathbb{E}[Y_\alpha|X_{i-2},X_{i+\ell+t+1}]-\mathbb{E}(Y_\alpha)|\leq\left(\sum_{x,y\in\mathcal{A}}\Delta_{x,y}\right)\left(\xi^t+\xi^{2t}\right) \tag{2.40}$$

comme :

$$\sum_{x,y\in\mathcal{A}}\Delta_{x,y}=\sum_{x,y\in\mathcal{A}}\sum_{\substack{b,a_1,\ldots,a_d,c\in\mathcal{A}\\ b\neq a_d}}\mu(x)\pi(x,b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\pi\left(a_r\,\mathbf{1}_{\{r\neq 0\}}+a_d\,\mathbf{1}_{\{r=0\}},c\right)$$
$$\times\,\pi(c,y)$$
$$=\sum_{c,y\in\mathcal{A}}\left(\sum_{\substack{x,b,a_1,\ldots,a_d\in\mathcal{A}\\ b\neq a_d}}\mu(x)\pi(x,b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\right)\pi\left(a_r\,\mathbf{1}_{\{r\neq 0\}}+a_d\,\mathbf{1}_{\{r=0\}},c\right)$$
$$\times\,\pi(c,y)$$

Nous avons d'après les propriétés des chaînes de Markov, d'une part :

$$\sum_{\substack{x,b,a_1,\ldots,a_d\in\mathcal{A}\\ b\neq a_d}}\mu(x)\pi(x,b)\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\leq 1$$

et d'autre part :

$$\begin{aligned}\sum_{x,y\in\mathcal{A}} \Delta_{x,y} &\leq \sum_{c,y\in\mathcal{A}} \pi\left(a_r\,\mathbf{1}_{\{r\neq 0\}} + a_d\,\mathbf{1}_{\{r=0\}}, c\right)\pi(c,y)\\ &= 1\end{aligned}$$

en remplaçant ces sommes par leurs majorations dans (2.40), nous obtenons :

$$|\mathbb{E}[Y_\alpha|X_{i-2}, X_{i+\ell+t+1}] - \mathbb{E}(Y_\alpha)| \leq \xi^t + \xi^{2t} \tag{2.41}$$

il s'en suit en remplaçant de nouveau (2.41) dans (2.32) :

$$b_3 \leq \sum_{\alpha\in\mathcal{I}} \left(\xi^t + \xi^{2t}\right) \tag{2.42}$$

Comme il y a une lettre qui sépare le mot W_α du mot W_β, la position i a au plus $n-t-2$ possibilités, et ℓ au plus $t-1$ positions (car c'est la longueur minimale du mot W_β), la majoration dans (2.42) est exprimée par :

$$\begin{aligned}b_3 &\leq \sum_{i=1}^{n-t-2}\sum_{\ell=1}^{t-1}\left(\xi^t + \xi^{2t}\right)\\ &= (n-t-2)(t-1)\left(\xi^t + \xi^{2t}\right)\\ &= n(t-1)\left(1 - \frac{t}{n} - \frac{2}{n}\right)\left(\xi^t + \xi^{2t}\right)\end{aligned}$$

en majorant $t-1$ par t, nous obtenons pour $t = o(n)$:

$$\begin{aligned}b_3 &\leq nt\left(\xi^t + \xi^{2t}\right)\\ &= \left(\frac{t}{n^3}\right)\left(n^2\xi^t\right)^2 + \left(\frac{t}{n}\right)n^2\xi^t\end{aligned}$$

C.Q.F.D

2.3.4 Preuve du théorème 2.2.1

La démonstration du théorème 2.2.1 résulte directement de l'addition membres à membres des majorations respectives des quantités b_1, b_2 et b_3 données dans la proposition 2.3.1, la proposition 2.3.2 et la proposition 2.3.3, ainsi :

$$\begin{aligned}b_1 + b_2 + b_3 &\leq 3\left(\frac{t}{n}\right)^2\left(n^2\,\xi^t\right)^2 + 24\left(\frac{t}{n}\right)^3\left(n^2\,\xi^t\right)^2 + \left(\frac{4C(\xi)}{n}\right)n^2\,\xi^t + \left(\frac{t}{n^3}\right)\left(n^2\xi^t\right)^2\\ &\quad + \left(\frac{t}{n}\right)n^2\xi^t\\ &= \left(\frac{24t^3+t}{n^3} + \frac{3t^2}{n^2}\right)\left(n^2\,\xi^t\right)^2 + \left(\frac{t+4C(\xi)}{n}\right)n^2\,\xi^t\end{aligned}$$

en posant :

$$\varphi(t,n) = \frac{24t^3+t}{n^3} + \frac{3t^2}{n^2} \qquad \text{et} \qquad \psi(t,n) = \frac{t+4C(\xi)}{n}$$

nous obtenons alors la majoration suivante de la distance en variation totale $d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right)$:

$$d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right) \leq 2\left(\varphi(t,n)\left(n^2\,\xi^t\right)^2 + \psi(t,n)\,n^2\,\xi^t\right)$$

Pour $n^2\xi^t = O(1)$, nous savons d'après le corollaire 1.3.1 (chapitre 1) que $t = O\left(\log_{1/\xi}(n)\right)$, l'expression donnant la majoration de la distance en variation totale $d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right)$ devient :

$$\begin{aligned} d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right) &\leq \frac{O\left(\log_{1/\xi}(n)\right)^3 + O\left(\log_{1/\xi}(n)\right)}{n^3} + \frac{O\left(\log_{1/\xi}(n)\right) + 4C(\xi)}{n} \\ &= O\left(\frac{\left(\log_{1/\xi}(n)\right)^3 + \log_{1/\xi}(n)}{n^3} + \frac{\log_{1/\xi}(n)+1}{n}\right) \end{aligned}$$

comme :

$$\lim_{n\to+\infty} \frac{\left(\log_{1/\xi}(n)\right)^3 + \log_{1/\xi}(n)}{n^3} + \frac{\log_{1/\xi}(n)+1}{n} = 0$$

alors $d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right)$ tend vers 0 quand n tend vers $+\infty$. Donc pour de très longues séquences, la loi de probabilité $\mathcal{L}(N_t)$ de N_t est approximée par la loi de Poisson $\mathcal{P}_\lambda$ où le paramètre λ est donné dans (1.10) .

2.4 Application à la significativité statistique

2.4.1 Significativité statistique

La significativité statistique a pour objet de discerner entre un événement qui est dû au hasard et un événement qui ne l'est pas. Nous dirons qu'un événement est :

- *non significatif* si sa réalisation est dû au hasard,
- *significatif* s'il a une probabilité (aussi faible qu'elle soit) de se réaliser, un tel événement est dit *exceptionnel*.

L'étude de la significativité statistique nécessite de calculer la p-value, elle est donnée par :

$$p = \mathbb{P}\left(X \geq x\right)$$

où X est la variable aléatoire utilisée pour décrire l'événement considéré, c'est donc une fonction de la suite de variables générées aléatoirement, et x est déterminé à partir des valeurs observées du phénomène étudié.

Dans notre modèle, il s'agit d'étudier la significativité statistique du nombre de répétitions chevauchantes maximales à gauche N_t^{obs} dans la séquence observée, c'est à dire de le comparer au nombre de répétitions chevauchantes maximales à gauche N_t, qui aurait pu se réaliser si la séquence est générée aléatoirement, autrement dit, voir si l'événement $\left\{N_t \geq N_t^{obs}\right\}$ est dû ou non au hasard. La p-value dans ce cas, est donnée par :

$$\begin{aligned} p &= \mathbb{P}\left(N_t \geq N_t^{obs}\right) \\ &= 1 - \mathbb{P}\left(N_t < N_t^{obs}\right) \\ &= 1 - \sum_{k=1}^{N_t^{obs}-1} \mathbb{P}\left(N_t = k\right) \end{aligned}$$

La loi de probabilité de N_t étant approximée d'après le théorème 2.2.1 par la loi de Poisson $\mathcal{P}_\lambda$, où le paramètre λ donné dans (1.10), nous pouvons alors obtenir une approximation de la p-value, donnée par :

$$p \approx 1 - \sum_{k=1}^{N_t^{obs}-1} \lambda^k \frac{e^{-\lambda}}{k!} \tag{2.43}$$

2.4.2 Calcul pratique de l'approximation de la p-value

La formule (2.43) nécessite pour avoir une valeur de la p-value de connaître la valeur de λ donnée dans (1.10). Ce qui revient à estimer λ, et à montrer aussi que l'approximation de la loi de probabilité de N_t par la loi de Poisson ayant pour paramètre l'estimateur de λ, reste valide.

Théorème 2.4.1. *Soit $\widehat{\lambda}$ l'estimateur de λ, sous les mêmes conditions du théorème 2.2.1, si $n^2\xi^t = O(1)$, alors :*

$$d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_{\widehat{\lambda}}\right) = o(1)$$

Démonstration. L'inégalité triangulaire appliquée à la distance en variation totale d_{VT} donne :

$$d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_{\widehat{\lambda}}\right) \leq d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right) + d_{VT}\left(\mathcal{P}_\lambda, \mathcal{P}_{\widehat{\lambda}}\right) \tag{2.44}$$

D'après le théorème 2.2.1, il reste à montrer que pour $n^2\xi^t = O(1)$:

$$d_{VT}\left(\mathcal{P}_\lambda, \mathcal{P}_{\widehat{\lambda}}\right) = o(1) \tag{2.45}$$

Nous avons d'après P.S.Rusakin (2004) [10] :

$$d_{VT}\left(\mathcal{P}_\lambda, \mathcal{P}_{\widehat{\lambda}}\right) \leq \left|\widehat{\lambda} - \lambda\right| \tag{2.46}$$

Il vient de la relation (1.10), que déterminer l'estimateur $\widehat{\lambda}$ de λ revient à déterminer les estimateurs de la loi initiale μ et de chacune des probabilités de transition π. Notons par $\widehat{\mu}$ et $\widehat{\pi}$ respectivement les estimateurs du maximum de vraisemblance de μ et π, elles sont d'après la loi du logarithme itéré appliquée aux chaînes de Markov (R.Senoussi (1990) [13]), telles que pour tout a et b dans $\mathcal{A}$:

$$\widehat{\mu}(b) = \mu(b) + O\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right) \quad \text{p.s} \qquad \text{et} \qquad \widehat{\pi}(a,b) = \pi(a,b) + O\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right) \quad \text{p.s}$$

en remplaçant μ et π respectivement par $\widehat{\mu}$ et $\widehat{\pi}$ dans (1.10), nous obtenons en posant aussi $\varepsilon_n = O\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right)$:

$$\begin{aligned}\widehat{\lambda} \;=\; & \sum_{\ell=1}^{t-1} \sum_{a_1,\ldots,a_d \in \mathcal{A}} (\mu(a_1) + \varepsilon_n) \left(\prod_{j=1}^{d-1} (\pi(a_j, a_{j+1}) + \varepsilon_n)\right) \widehat{\Gamma} + \sum_{\ell=1}^{t-1} (n - t - \ell) \\ & \sum_{\substack{b,a_1,\ldots,a_d \in \mathcal{A} \\ b \neq a_d}} (\mu(b) + \varepsilon_n)\,(\pi(b, a_1) + \varepsilon_n) \left(\prod_{j=1}^{d-1} (\pi(a_j, a_{j+1}) + \varepsilon_n)\right) \widehat{\Gamma}\end{aligned} \tag{2.47}$$

avec :

$$\widehat{\Gamma} = \left(\prod_{j=1}^{d-1}\left(\pi(a_j,a_{j+1})+\varepsilon_n\right)\right)^{q-1}\left(\pi(a_d,a_1)+\varepsilon_n\right)^{q-1} \times \left(\left(\pi(a_d,a_1)+\varepsilon_n\right)\prod_{j=1}^{r-1}\left(\pi(a_j,a_{j+1})+\varepsilon_n\right)\right)^{\mathbf{1}_{\{r\neq 0\}}}$$

en développant le produit (2.47) en gardant uniquement λ et les termes d'ordre ε_n, nous obtenons l'écriture de $\widehat{\lambda}$ au voisinage de λ, autrement dit, un développement limité de $\widehat{\lambda}$ au voisinage de λ, ainsi :

$$\begin{aligned}\widehat{\lambda} &= \lambda + \varepsilon_n\left(\sum_{\ell=1}^{t-1}\sum_{a_2,\ldots,a_d\in\mathcal{A}}\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\Gamma + \sum_{\ell=1}^{t-1}(n-t-\ell)\sum_{a_1,\ldots,a_d\in\mathcal{A}}\pi(b,a_1)\right.\\ &\left.\times\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)\Gamma\right) \qquad (2.48)\end{aligned}$$

où Γ est donné par (1.8), et qui d'après la démonstration du lemme 1.3.2 (chapitre 1) est majoré par ξ^t quand chacune de ses probabilités de transition est majorée par ξ $(0<\xi<1)$, et comme :

$$\sum_{a_2,\ldots,a_d\in\mathcal{A}}\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)=1 \qquad \text{et} \qquad \sum_{a_1,\ldots,a_d\in\mathcal{A}}\pi(b,a_1)\left(\prod_{j=1}^{d-1}\pi(a_j,a_{j+1})\right)=1$$

il s'en suit alors en remplaçant dans (2.48), après avoir utilisé des calculs similaires à ceux de la démonstration de la proposition 1.3.1 (chapitre 1), pour $t=o(n)$, qu'au voisinage de λ :

$$\widehat{\lambda} = \lambda + O\left(n^2\xi^t\,\varepsilon_n\right)$$

ce qui donne pour $n^2\xi^t = O(1)$:

$$\widehat{\lambda} = \lambda + O\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right)$$

en remplaçant par la suite dans (2.46), nous retrouvons (2.45). En remplaçant de nouveau (2.45) dans l'inégalité triangulaire (2.44), en appliquant en plus le théorème 2.2.1, nous déduisons le résultat. C.Q.F.D

Chapitre 3

Généralisation aux chaînes de Markov d'ordre m

Dans le chapitre 2, nous avons montré que sous les hypothèses de la proposition 1.3.1 et du corollaire 1.3.1 (chapitre 1), pour lesquels le paramètre λ est borné sur $]0, +\infty[$, que la loi de probabilité de N_t est approximée par la loi de Poisson $\mathcal{P}_\lambda$ avec λ donnée dans (1.10). Dans ce chapitre nous donnons une généralisation du théorème 2.2.1 dans le cas où la séquence S est extraite d'une suite de variables aléatoires $(X_n)_n$ modélisée par une chaîne de Markov d'ordre m où $m \geq 1$.

Sommaire

3.1 Le Modèle Mm

Dans la suite une de chaîne de markov d'ordre m ($m \geq 1$) est dite chaîne de Markov Mm, que nous appelons aussi modèle Mm.

3.1.1 Modélisation d'une séquence à l'aide d'une chaîne Mm

Définition 3.1.1 (Chaîne de Markov d'ordre m). *Soit $(X_n)_n$ une suite de variables aléatoires à valeurs dans un espace d'état $\mathcal{A}$. On dit que $(X_n)_n$ est une* chaîne de Markov d'ordre m *si pour tout $a_1, \ldots, a_n$ dans $\mathcal{A}$:*

$$\mathbb{P}\left(X_n = a_n \,/\, X_1 = a_1, \ldots, X_{n-1} = a_{n-1}\right) = \mathbb{P}\left(X_n = a_n \,/\, X_{n-m} = a_{n-m}, \ldots, X_{n-1} = a_{n-1}\right)$$

D'après la définition 3.1.1, la connaissance de la valeur du m-uplet $(X_{n-m}, \ldots, X_{n-1})$ suffit pour déterminer la probabilité pour la variable aléatoire X_n de prendre la valeur a_n, la proba-

bilité $\mathbb{P}\left(X_n = a_n \,/\, X_{n-m} = a_{n-m}, \ldots, X_{n-1} = a_{n-1}\right)$ est appelée probabilité de transition de $a_{n-m} \cdots a_{n-1}$ à a_n, elle est notée $\pi(a_{n-m} \cdots a_{n-1}, a_n)$.

La loi initiale μ est définie par :

$$\mu(a_1, \ldots, a_m) = \mathbb{P}\left(X_1 = a_1, \ldots, X_m = a_m\right)$$

Dans ce chapitre les séquences sont modélisées par une chaîne de Markov Mm. Soit donc $S = X_1 X_2 \cdots X_n$ une séquence de longueur n générée par une chaîne de Markov Mm à valeur sur un alphabet fini $\mathcal{A}$. Soit $\Pi = (\pi(A, b))_{A \in \mathcal{A}^m, b \in \mathcal{A}}$ la matrice de probabilités de transition définies par :

$$\forall\, A = a_1 \cdots a_m \in \mathcal{A}^m \quad \forall\, b \in \mathbb{A} \qquad \pi(A, b) = \mathbb{P}\left(X_i = b \,/\, X_{i-m} = a_1, \ldots, X_{i-1} = a_m\right)$$

et μ la loi initiale, que nous supposons stationnaire, définie par :

$$\forall\, i = 1, \ldots, n - m + 1 \qquad \mu(A) = \mathbb{P}(X_i = a_1, \ldots, X_{i+m-1} = a_m)$$

Pour tout $i = 1, \ldots, n - m + 1$, on note par :

$$\forall\, i = 1, \ldots, n - m + 1 \qquad X_i^{(m)} = X_i X_{i+1} \cdots X_{i+m-1} \tag{3.1}$$

un bloc de m variables aléatoires, il s'en suit de (3.1) que :

(a) le bloc $X_i^{(m)}$ (où $i = 1, \ldots, n - m + 1$) est une variable aléatoire à valeurs dans $\mathcal{A}^m$;

(b) la séquence S est réécrite sous la forme d'une séquence $S^{(m)}$, que nous désignons par :

$$S^{(m)} = X_1^{(m)} \cdots X_{n-m+1}^{(m)}$$

ayant pour longueur $n - m + 1$.

Il en découle de (a) et (b) que $X_{i+1} \cdots X_{i+m-1}$ appartiennent aux deux blocs de variables $X_i^{(m)}$ et $X_{i+1}^{(m)}$, la transition de $X_i^{(m)}$ à $X_{i+1}^{(m)}$ correspond donc à la transition de la variable aléatoire X_{i+m-1} à X_{i+m}. Nous définissons alors pour tout $A = a_1 \cdots a_m$ et $B = b_1 \cdots b_m$ dans $\mathcal{A}^m$, la probabilité de transition π_m de A à B par :

$$\pi_m(A, B) = \begin{cases} \pi(A, b_m) & \text{si} \quad a_2 \cdots a_m = b_1 \cdots b_{m-1} \\ \\ 0 & \text{sinon} \end{cases}$$

Si $w_i = X_i \cdots X_{i+t-1}$ est un mot de longueur t dans la séquence S, dans $S^{(m)}$ c'est un mot $w_i^{(m)} = X_i^{(m)} \cdots X_{i+t-m}^{(m)}$ de longueur $t - m + 1$. En une position $\alpha = (i, j)$ de $\mathcal{I}$, une répétition est donc chevauchante, si au moins les m premières lettres de $w_j^{(m)}$ recouvrent les m dernières de $w_i^{(m)}$, ce qui ce traduit par :

$$\ell = |i - j| \leq t - m$$

De même que dans la sous-section 1.1.1 (chapitre 1), nous supposons par symétrie que $i < j$, la position j alors est telle que $j = i + \ell$. Une répétition chevauchante maximale à gauche à la position α dans la séquence $S^{(m)}$ est un mot $\mathrm{W}_\alpha^{(m)}$ de longueur $t + \ell - m + 1$. Comme la longueur de la séquence $S^{(m)}$ est $n-m+1$, alors i a $(n-m+1)-(t+\ell-m+1)+1 = n-t-\ell+1$ positions, et l'ensemble des positions α des répétitions chevauchantes maximales à gauche dans la séquence $S^{(m)}$ est donc :

$$\mathcal{I}^{(m)} = \{\alpha = (i, i + \ell);\ \ell = 1, \ldots, t - m;\ i = 1, \ldots, n - t - \ell + 1\}$$

Exemple 3.1.1. Considérons la séquence « $S =$ `tacttgacttgacttgacttgactg` » de l'exemple 1.1.3 (chapitre 1), S est de longueur $n = 25$. Si S est extraite d'une chaîne de Markov d'ordre $m = 3$, prenant ses valeurs dans $\mathcal{A}$, en posant :

$$\texttt{T=tac}, \quad \texttt{A=act}, \quad \texttt{B=ctt}, \quad \texttt{C=ttg}, \quad \texttt{D=tga} \quad \text{et} \quad \texttt{E=gac}$$

elle devient alors en la réécrivant avec ces blocs de 3 lettres, la séquence :

$$S^{(3)} = \texttt{T A B C D E A B C D E A B C D E A B C D E A F}$$

à valeurs dans $\mathcal{A}^3$, ayant pour longueur $n - m + 1 = 25 - 3 + 1 = 23$.

En une position $\alpha = (i, i + \ell)$ de $\mathcal{I}^{(m)}$, une répétition chevauchante et maximale à gauche dans $S^{(m)}$ est définie par :

$$X_{i-1}^{(m)} \neq X_{i+\ell-1}^{(m)} \quad \text{et} \quad w_i^{(m)} = w_{i+\ell}^{(m)} \tag{3.2}$$

Exemple 3.1.2. Reprenons la séquence $S^{(3)}$ de l'exemple 3.1.1 et l'exemple 1.1.3 (chapitre 1), à la position $(2, 12)$ le mot $\mathrm{W}_{(2,12)} =$ `acttgacttgacttgacttgact`, qui est la répétition du mot $w_2 =$ `acttgacttgact` dans la séquence S, devient dans la séquence $S^{(3)}$ le mot :

$$\mathrm{W}_{(2,12)}^{(3)} = \texttt{ABCDEABCDEABCDEABCDEA}$$

qui est la répétition du mot :

$$w_2^{(3)} = \texttt{ABCDEABCDEA}$$

de longueur $t - 3 + 1 = 11$ dans la séquence $S^{(3)}$. Le mot $\mathrm{W}_{(2,12)}^{(3)}$ a donc pour longueur $t + \ell - 3 + 1 = 21$ ($\ell = 10$) (figure 3.1),

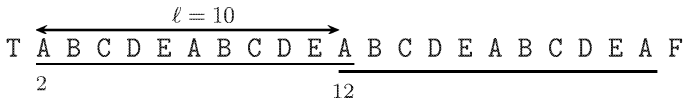

FIG. 3.1 – Le mot $\mathrm{W}_{(2,12)}^{(3)}$ est la répétition du mot $w_2^{(3)}$ en $(2, 12)$.

Lemme 3.1.1. *La maximalité à gauche de la répétition en α dans la séquence $S^{(m)}$ donnée dans* (3.2)*, est équivalente à :*

$$X_{i-1} \neq X_{i+\ell-1}$$

Démonstration. Nous avons d'après (3.2) :

$$w_i^{(m)} = w_{i+\ell}^{(m)} \Leftrightarrow \begin{cases} X_i^{(m)} & = X_{i+\ell}^{(m)} \\ & \vdots \\ X_t^{(m)} & = X_{i+\ell+t-m}^{(m)} \end{cases} \tag{3.3}$$

comme d'après (3.1) :

$$X_{i-1}^{(m)} = X_{i-1}X_i \cdots X_{i+m-2} \qquad \text{et} \qquad X_{i+\ell-1}^{(m)} = X_{i+\ell-1}X_{i+\ell} \cdots X_{i+\ell+m-1}$$

alors les blocs de variables aléatoires $X_i \cdots X_{i+m-2}$ et $X_{i+\ell} \cdots X_{i+\ell+m-1}$ sont d'après (3.3) égaux, ainsi seules les premières lettres X_{i-1} et $X_{i+\ell-1}$ des blocs $X_{i-1}^{(m)}$ et $X_{i+\ell-1}^{(m)}$ sont différentes, d'où le résultat. C.Q.F.D

Exemple 3.1.3. Dans l'exemple 3.1.2, nous avons $X_1^{(3)} = \mathtt{T}$ et $X_{11}^{(3)} = \mathtt{E}$, et on a bien :

$$\mathtt{T} \neq \mathtt{E} \Leftrightarrow \mathtt{tac} \neq \mathtt{gac} \Leftrightarrow \mathtt{t} \neq \mathtt{g}$$

Pour tout $\alpha = (i, i+\ell)$ dans $\mathcal{I}^{(m)}$, on définit la variable aléatoire indicatrice $Y_\alpha^{(m)}$ qui compte les répétitions chevauchantes maximales à gauche dans la séquence $S^{(m)}$ par :

$$Y_\alpha^{(m)} = \begin{cases} \mathbf{1}_{\left\{w_1^{(m)} = w_{1+\ell}^{(m)}\right\}} & \text{si} \quad i = 1 \\ \mathbf{1}_{\left\{X_{i-1}^{(m)} \neq X_{i+\ell-1}^{(m)},\, w_i^{(m)} = w_{i+\ell}^{(m)}\right\}} & \text{si} \quad i > 1 \end{cases} \tag{3.4}$$

Lemme 3.1.2. *Les variables aléatoires indicatrices $Y_\alpha^{(m)}$ et Y_α, qui comptent les répétitions chevauchantes maximales à gauche respectivement dans $S^{(m)}$ et S, sont égales.*

Démonstration. D'après la définition de $Y_\alpha^{(m)}$ dans (3.4), nous avons à considérer deux cas, celui où $i = 1$ et celui où $i > 1$.

1. **Pour $i = 1$.** Nous avons en vertu de (3.1), (3.3) et (3.4) :

$$\begin{aligned} Y_\alpha^{(m)} = 1 &\Leftrightarrow \mathbf{1}_{\left\{w_1^{(m)} = w_{1+\ell}^{(m)}\right\}} = 1 \\ &\Leftrightarrow X_1^{(m)} \cdots X_t^{(m)} = X_{1+\ell}^{(m)} \cdots X_{t+\ell}^{(m)} \\ &\Leftrightarrow \begin{cases} X_1 \cdots X_m & = X_{1+\ell} \cdots X_{1+\ell+m} \\ & \vdots \\ X_{t-m} \cdots X_t & = X_{t+\ell+t-m} \cdots X_{t-\ell} \end{cases} \\ &\Leftrightarrow X_1 \cdots X_t = X_{1+\ell} \cdots X_{t+\ell} \\ &\Leftrightarrow \mathbf{1}_{\{w_1 = w_{1+\ell}\}} = 1 \\ &\Leftrightarrow Y_\alpha = 1 \end{aligned}$$

2. **Pour $i > 1$.** De même que dans 1. ci-dessus, nous avons aussi en appliquant en plus le lemme 3.1.1 :

$$\begin{aligned} Y_\alpha^{(m)} = 1 &\Leftrightarrow \mathbf{1}_{\left\{X_{i-1}^{(m)} \neq X_{i+\ell-1}^{(m)},\, w_i^{(m)} = w_{i+\ell}^{(m)}\right\}} = 1 \\ &\Leftrightarrow \begin{cases} X_{i-1}^{(m)} \neq X_{i+\ell-1}^{(m)} \\ X_i^{(m)} \cdots X_{i+t-1}^{(m)} = X_{i+\ell}^{(m)} \cdots X_{i+t+\ell-1}^{(m)} \end{cases} \\ &\Leftrightarrow \begin{cases} X_{i-1} \neq X_{i+\ell-1} \\ X_i \cdots X_{i+m-1} & = X_{i+\ell} \cdots X_{i+\ell+m-1} \\ & \vdots \\ X_{i+t-m-1} \cdots X_{i+t-1} & = X_{i+t+\ell-m-1} \cdots X_{i+t-\ell-1} \end{cases} \\ &\Leftrightarrow \begin{cases} X_{i-1} \neq X_{i+\ell-1} \\ X_i \cdots X_{i+t-1} = X_{i+\ell} \cdots X_{i+t+\ell-1} \end{cases} \\ &\Leftrightarrow \mathbf{1}_{\{X_{i-1} \neq X_{i+\ell-1},\, w_i = w_{i+\ell}\}} = 1 \\ &\Leftrightarrow Y_\alpha = 1 \end{aligned}$$

Nous déduisons de 1. et 2., que $Y_\alpha^{(m)} = Y_\alpha$. C.Q.F.D

3.1.2 Nombre moyen de répétitions $\lambda^{(m)}$ et ordre de grandeur

D'après (a) et (b), ainsi que le lemme 3.1.1 et le lemme 3.1.2, la séquence S de longueur n générée par une chaîne de Markov d'ordre m à valeurs dans $\mathcal{A}$. Écrite sous la forme d'une séquence $S^{(m)}$ de longueur $n-m+1$, elle est considérée comme une séquence générée par une chaîne de Markov d'ordre 1 à valeurs dans $\mathcal{A}^m$.

Soit $N_{t-m+1}^{(m)}$ le nombre de répétitions chevauchantes maximales à gauche dans la séquence $S^{(m)}$. Soit $\lambda^{(m)}$ le nombre moyen de répétitions chevauchantes maximales à gauche dans la séquence $S^{(m)}$, nous avons :

$$N_{t-m+1}^{(m)} = \sum_{\alpha \in \mathcal{I}^{(m)}} Y_\alpha^{(m)} \qquad \text{et} \qquad \lambda^{(m)} = \mathbb{E}\left(N_{t-m+1}^{(m)}\right)$$

pour ℓ fixé, i a $(n-m+1)-(t+\ell-m+1)+1 = n-t-\ell+1$ positions, ainsi :

$$N_{t-m+1}^{(m)} = \sum_{\ell=1}^{t-m} \sum_{i=1}^{n-t-\ell+1} Y_\alpha^{(m)}$$

Remarque 3.1.1. Notons qu'en vertu du lemme 3.1.2, le nombre $N_{t-m+1}^{(m)}$ est tel que :

$$N_{t-m+1}^{(m)} = N_t$$

où N_t est le nombre de répétitions chevauchantes maximales à gauche dans S.

Le nombre moyen $\lambda^{(m)}$ de répétitions chevauchantes maximales à gauche dans la séquence $S^{(m)}$ est donné par :

$$\lambda^{(m)} = \sum_{\ell=1}^{t-m} \sum_{i=1}^{n-t-\ell+1} \mathbb{E}\left(Y_\alpha^{(m)}\right)$$

Lemme 3.1.3. *Soient $A_1, \ldots, A_d$ les valeurs prises les $t+\ell-m+1$ blocs de lettres de $\mathrm{W}_\alpha^{(m)}$, le nombre moyen $\lambda^{(m)}$ de répétitions chevauchantes maximales à gauche dans la séquence $S^{(m)}$, est tel que :*

$$\lambda^{(m)} = \lambda_1^{(m)} + \lambda_2^{(m)} \tag{3.5}$$

où :

$$\begin{aligned} \lambda_1^{(m)} &= \sum_{\ell=1}^{t-m} \sum_{A_1,\ldots,A_d \in \mathcal{A}^m} \mu(A_1) \left(\prod_{j=1}^{d-1} \pi_m(A_j, A_{j+1})\right)^q (\pi_m(A_d, A_1))^{q-1} \\ &\quad \times \left(\pi_m(A_d, A_1) \prod_{j=1}^{r-1} \pi_m(A_j, A_{j+1})\right)^{\mathbf{1}_{\{r\neq 0\}}} \end{aligned} \tag{3.6}$$

et

$$\begin{aligned} \lambda_2^{(m)} &= \sum_{\ell=1}^{t-m} (n-t-\ell) \sum_{\substack{B,A_1,\ldots,A_d \in \mathcal{A}^m \\ B \neq A_d}} \mu(B)\, \pi_m(B, A_1) \left(\prod_{j=1}^{d-1} \pi_m(A_j, A_{j+1})\right)^q \\ &\quad \times (\pi_m(A_d, A_1))^{q-1} \left(\pi_m(A_d, A_1) \prod_{j=1}^{r-1} \pi_m(A_j, A_{j+1})\right)^{\mathbf{1}_{\{r\neq 0\}}} \end{aligned} \tag{3.7}$$

Démonstration. D'après la linéarité de l'espérance mathématique appliquée à $\lambda^{(m)}$, nous avons :

$$\lambda^{(m)} = \sum_{\ell=1}^{t-m} \mathbb{P}\left(w_1^{(m)} = w_{1+\ell}^{(m)}\right) + \sum_{\ell=1}^{t-m} \sum_{i>1}^{n-t-\ell+1} \mathbb{P}\left(X_{i-1}^{(m)} \neq X_{i+\ell-1}^{(m)};\, w_i^{(m)} = w_{i+\ell}^{(m)}\right)$$

Posons :

$$\lambda_1^{(m)} = \sum_{\ell=1}^{t-m} \mathbb{P}\left(w_1^{(m)} = w_{1+\ell}^{(m)}\right) \tag{3.8}$$

et

$$\lambda_2^{(m)} = \sum_{\ell=1}^{t-m} \sum_{i>1}^{n-t-\ell+1} \mathbb{P}\left(X_{i-1}^{(m)} \neq X_{i+\ell-1}^{(m)};\, w_i^{(m)} = w_{i+\ell}^{(m)}\right) \tag{3.9}$$

en utilisant le même raisonnement que dans la sous-section 1.3.1 (chapitre 1), d'après la caractérisation des répétitions chevauchantes maximales à gauche, le mot $\mathrm{W}_\alpha^{(m)}$ est la répétition des d ($d \leq \ell$) premiers blocs de m lettres de $w_i^{(m)}$ où d est tel que :

$$\ell = kd\ (k \in \mathbb{N}^*) \qquad \text{et} \qquad t + \ell - m + 1 = qd + r \qquad (0 \leq r < d)$$

Soient $A_1, \ldots A_d$ les valeurs prises par les lettres du mot $\mathrm{W}_\alpha^{(m)}$, nous avons donc :

1. **pour** $i = 1$ **:**

$$\begin{aligned}\mathbb{P}\left(w_1^{(m)} = w_{1+\ell}^{(m)}\right) \;=\; & \sum_{A_1,\ldots,A_d \in \mathcal{A}^m} \mu(A_1) \left(\prod_{j=1}^{d-1} \pi_m(A_j, A_{j+1})\right)^q (\pi_m(A_d, A_1))^{q-1} \\ & \times \left(\pi_m(A_d, A_1) \prod_{j=1}^{r-1} \pi_m(A_j, A_{j+1})\right)^{\mathbb{I}_{\{r\neq 0\}}}\end{aligned} \tag{3.10}$$

2. **pour** $i > 1$ **:**

$$\begin{aligned}&\mathbb{P}\left(X_{i-1}^{(m)} \neq X_{i+\ell-1}^{(m)};\, w_i^{(m)} = w_{i+\ell}^{(m)}\right) \\ &\quad = \sum_{\substack{B,A_1,\ldots,A_d \in \mathcal{A}^m \\ B \neq A_d}} \mu(B)\, \pi_m(B, A_1) \left(\prod_{j=1}^{d-1} \pi_m(A_j, A_{j+1})\right)^q (\pi_m(A_d, A_1))^{q-1} \\ &\qquad\qquad \times \left(\pi_m(A_d, A_1) \prod_{j=1}^{r-1} \pi_m(A_j, A_{j+1})\right)^{\mathbb{I}_{\{r\neq 0\}}}\end{aligned} \tag{3.11}$$

nous obtenons en remplaçant (3.10) dans (3.8), et (3.11) dans (3.9) tout en tenant compte pour ce cas de la condition de maximalité à gauche, respectivement (3.6) et (3.7), et en additionnant membres à membres (3.6) et (3.7), nous déduisons (3.5). C.Q.F.D

Posons pour tout A et B dans $\mathcal{A}^m$:

$$\xi = \max_{A,B \in \mathcal{A}^m} \pi_m(A, B) \qquad (0 < \xi < 1)$$

Proposition 3.1.1. *Pour $t-m+1=o(n)$, on a :*

$$\lambda^{(m)} = O\left(n^2\,\xi^{t-m+1}\right)$$

Si $n^2\,\xi^{t-m+1}=O(1)$, alors $\lambda^{(m)}$ est borné sur $]0,+\infty[$.

Démonstration. En utilisant un raisonnement analogue à celui utilisé dans les démonstrations de la section 1.3.2 (chapitre 1), nous obtenons en majorant chacune des probabilités de transition dans (3.10) et (3.11) par ξ $(0<\xi<1)$:

$$\mathbb{P}\left(w_1^{(m)}=w_{1+\ell}^{(m)}\right)\leq\xi^{t-m+1}\qquad\text{et}\qquad\mathbb{P}\left(X_{i-1}^{(m)}\neq X_{i+\ell-1}^{(m)};\,w_i^{(m)}=w_{i+\ell}^{(m)}\right)\leq\xi^{t-m+1}$$

ces majorations remplacées dans (3.6) et (3.7), donne :

$$\lambda_1^{(m)}\leq(t-m)\,\xi^{t-m+1}\qquad\text{et}\qquad\lambda_2^{(m)}\leq(t-m)\left(n-m+1-\frac{3}{2}(t-m+1)\right)\xi^{t-m+1}$$

il en résulte en remplaçant de nouveau les majorations de $\lambda_1^{(m)}$ et $\lambda_2^{(m)}$ ci-dessus dans (3.5), tout en majorant $t-m$ par n, que pour $t-m+1=o(n)$:

$$\lambda^{(m)}\leq n^2\,\xi^{t-m+1}$$

et si $n^2\,\xi^{t-m+1}=O(1)$, alors $\lambda^{(m)}$ borné sur sur $]0,+\infty[$. C.Q.F.D

3.2 Théorème d'approximation pour le modèle Mm

Comme dans le chapitre 2, les quantités b_1, b_2 et b_3 dépendent du choix du voisinage des positions de répétitions. Nous commençons par définir la notion de voisinage de deux positions de deux répétitions dans la séquence $S^{(m)}$.

3.2.1 Choix du voisinage pour le modèle Mm

Soient $\alpha=(i,i+\ell)$ et $\beta=(i'+\ell')$ deux éléments de $\mathcal{I}^{(m)}$, positions de deux répétitions chevauchantes maximales à gauche de mots de longueur $t-m+1$.

Définition 3.2.1. *On dit que les positions α et β sont* non voisines *s'il existe au moins un* bloc *de m lettres qui sépare les mots $\mathrm{W}_\alpha^{(m)}$ et $\mathrm{W}_\beta^{(m)}$, et* voisines *sinon.*

D'après la définition 3.2.1, les positions α et β sont non voisines si les composantes de α et β sont telles que :

(a) si $i'-(i+\ell)>t-m+1$ si le mot $\mathrm{W}_\alpha^{(m)}$ occurre avant $\mathrm{W}_\beta^{(m)}$ (figure 3.2),

(b) si $i-(i'+\ell)'>t-m+1$ si le mot $\mathrm{W}_\alpha^{(m)}$ occurre après $\mathrm{W}_\beta^{(m)}$.

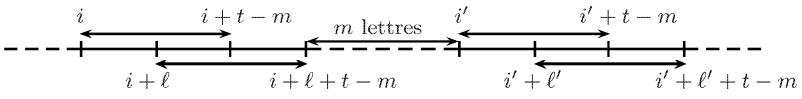

FIG. 3.2 – Les positions $\alpha=(i,i+\ell)$ et $\beta=(i',i'+\ell')$ sont non voisines.

Les quantités b_1, b_2 et b_3 sont définies et notées dans le modèle Mm par :

$$b_1^{(m)} = \sum_{\alpha \in \mathcal{I}^{(m)}} \sum_{\beta \in \mathrm{B}_\alpha^{(m)}} \mathbb{E}\left(Y_\alpha^{(m)}\right) \mathbb{E}\left(Y_\beta^{(m)}\right)$$

$$b_2^{(m)} = \sum_{\alpha \in \mathcal{I}^{(m)}} \sum_{\substack{\beta \in \mathrm{B}_\alpha^{(m)} \\ \beta \neq \alpha}} \mathbb{E}\left(Y_\alpha^{(m)} Y_\beta^{(m)}\right)$$

$$b_3^{(m)} = \sum_{\alpha \in \mathcal{I}^{(m)}} \mathbb{E}\left|\mathbb{E}\left(Y_\alpha^{(m)} - \mathbb{E}\left(Y_\alpha^{(m)}\right) | \sigma\left(Y_\beta^{(m)};\ \beta \notin \mathrm{B}_\alpha^{(m)}\right)\right)\right|$$

où $\mathrm{B}_\alpha^{(m)}$ est le voisinage de α, c'est à dire l'ensemble des éléments β de $\mathcal{I}^{(m)}$ voisins de α au sens de la définition 3.2.1.

3.2.2 Énoncé du théorème d'approximation pour le modèle Mm

D'après la remarque 3.1 et le lemme 3.1.2, majorer $d_{VT}\left(\mathcal{L}\left(N_t\right), \mathcal{P}_{\lambda^{(m)}}\right)$ où N_t est le nombre de répétitions chevauchantes maximales à gauche dans la séquence S, revient à majorer la borne supérieure de $d_{VT}\left(\mathcal{L}\left(N_{t-m+1}^{(m)}\right), \mathcal{P}_{\lambda^{(m)}}\right)$ où $N_{t-m+1}^{(m)}$ est le nombre de répétitions chevauchantes maximales à gauche dans la séquence $S^{(m)}$.

Théorème 3.2.1 (Corollaire du théorème 2.2.1). *Soient S une séquence générée par une chaîne de Markov d'ordre m à valeurs sur un alphabet fini $\mathcal{A}$, et N_t le nombre de répétitions chevauchantes maximales à gauche, et soit $\lambda^{(m)}$ le nombre moyen de répétitions chevauchantes maximales à gauche donné dans* (3.5). *Si $n^2\,\xi^{t-m+1} = O(1)$, alors :*

$$d_{VT}\left(\mathcal{L}\left(N_t\right), \mathcal{P}_{\lambda^{(m)}}\right) = o(1)$$

Démonstration. En réécrivant la séquence S extraite d'une chaîne de Markov d'ordre m en regroupant les variables aléatoires qui la composent en blocs de m variables consécutives, nous obtenons la séquence $S^{(m)}$ où ces blocs sont des variables aléatoires extraites d'une chaîne de Markov d'ordre 1 à valeurs dans $\mathcal{A}^{(m)}$. Par analogie à ce qui a été fait dans chapitre 2, pour les séquences modélisées par une chaîne de Markov d'ordre 1, cela revient à montrer que :

$$d_{VT}\left(\mathcal{L}\left(N_{t-m+1}^{(m)}\right), \mathcal{P}_{\lambda^{(m)}}\right) \leq 2\left(b_1^{(m)} + b_2^{(m)} + b_3^{(m)}\right)$$

En appliquant le théorème 2.2.1, nous avons bien que pour $n^2\,\xi^{t-m+1} = O(1)$:

$$d_{VT}\left(\mathcal{L}\left(N_{t-m+1}^{(m)}\right), \mathcal{P}_{\lambda^{(m)}}\right) = o(1) \tag{3.12}$$

et le résultat se déduit en remplaçant dans (3.12) d'après le lemme 3.1.2, le nombre $N_{t-m+1}^{(m)}$ dans la séquence $S^{(m)}$ par le nombre N_t dans la séquence S. C.Q.F.D

Ainsi si $n^2\,\xi^{t-m+1} = O(1)$, la distribution de probabilité du nombre de répétitions chevauchantes maximales à gauche dans une séquence générée par une chaîne de Markov d'ordre m, est aussi approximée par une loi de Poisson de paramètre $\lambda^{(m)}$, donné par (3.5).

Chapitre 4

Conclusion

L'analyse statistique du nombre N_t^{obs} de répétitions chevauchantes maximales à gauche de mots de longueur t, dans une séquence observée de longueur n donnée, que nous avons évoquée dans notre travail, consiste à étudier la significativité statistique du nombre N_t^{obs}. Il est nécessaire pour cela de calculer la p-value $p = \mathbb{P}\left(N_t \geq N_t^{obs}\right)$ où N_t représente le nombre de répétitions chevauchantes maximales à gauche dans une séquence générée aléatoirement, que l'on compare à la séquence observée. Ce calcul ne peut se faire sans la connaissance de la loi de probabilité exacte $\mathcal{L}(N_t)$ de N_t, ou du moins d'une approximation de $\mathcal{L}(N_t)$ par une loi de probabilité usuelle, si la loi de probabilité exacte ne peut être déterminée. Le travail développé dans cette thèse, a pour objectif de montrer que sous certaines conditions, la loi de probabilité $\mathcal{L}(N_t)$ de N_t est approximée par une loi de Poisson $\mathcal{P}_\lambda$ de paramètre λ, où $\lambda = \mathbb{E}(N_t)$ pour de très longues séquences, lorsque la séquence aléatoire est extraite d'une chaîne de Markov d'ordre 1, homogène et stationnaire à valeurs sur un alphabet fini. De cette approximation, celle de la p-value est alors déduite, ce qui permet donc de voir, si le nombre N_t^{obs} observé est statistiquement significatif ou non.

Nous avons commencé par donner la caractérisation d'une répétition chevauchante maximale à gauche, à partir de laquelle, nous avons déterminé l'expression de N_t, ainsi que celle du nombre moyen de ces répétitions $\lambda = \mathbb{E}(N_t)$. L'expression de λ est composée essentiellement de produit de probabilités de transition. En majorant ensuite chacune des probabilités de transition dans l'expression de λ par ξ $(0 < \xi < 1)$, nous avons montré que pour que $t = o(n)$, le paramètre λ est majoré par $n^2\xi^t$, et que si de plus $n^2\xi^t = O(1)$, alors λ est borné sur $]0, +\infty[$. Ce qui justifie par la suite, l'utilisation de λ comme paramètre de la loi de Poisson $\mathcal{P}_\lambda$ pour montrer l'approximation de $\mathcal{L}(N_t)$ par $\mathcal{P}_\lambda$.

La contribution que nous avons apportée dans ce travail, a été en un premier temps de donner des conditions suffisantes de maximalité à gauche pour les répétitions chevauchantes lorsqu'il y a recouvrement entre les répétitions chevauchantes, car dans ce cas la répétition qui recouvre l'autre n'est pas forcément maximale à gauche. Ensuite de montrer, en nous basant sur la méthode de Chen-Stein, que sous la condition $n^2\xi^t = O(1)$, la distance en variation totale $d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_\lambda\right)$ tend vers 0 pour des séquences très longues, autrement dit, la loi de probabilité de N_t est approximée par la loi de Poisson $\mathcal{P}_\lambda$. Nous avons montré aussi que si le paramètre λ est estimé par $\widehat{\lambda}$, la distance en variation totale $d_{VT}\left(\mathcal{L}(N_t), \mathcal{P}_{\widehat{\lambda}}\right)$ tend vers

0 pour de très longues séquences, dès que la condition $n^2\xi^t = O(1)$ est réalisée. Partant de cette approximation poissonnienne, nous avons présenté un calcul pratique permettant d'obtenir une valeur approximative de la p-value. Suite au résultat établi pour la modèlisation par une chaîne de Markov d'ordre 1, homogène et stationnaire des séquences aléatoires, nous avons montré en utilisant une modélisation par une chaîne de Markov d'ordre m ($m \geq 1$), homogène et stationnaire à valeurs sur un alphabet fini, que sous la condition $n^2\xi^{t-m+1} = O(1)$, l'approximation de la loi de probabilité $\mathcal{L}(N_t)$ du nombre de répétitions chevauchantes maximale à gauche, par une loi de Poisson de paramètre $\lambda^{(m)}$, déterminé pour ce modèle, reste encore valide.

Par ce travail, nous avons donc donné une extension et complété des résultats établis auparavant. D'une part, par R.Arratia *and al.* (1996) [3], concernant l'approximation poissonnienne de la loi de probabilité du nombre de répétitions maximales à gauche (chevauchantes et non chevauchantes) quand la séquence aléatoire est extraite d'une suite de variables aléatoires indépendantes et identiquement distribuées (i-i-d). Et d'autre part, par N.Touyyar *and al.* (2008) [15], concernant aussi l'approximation poissonnienne de la loi de probabilité du nombre de répétitions non chevauchantes maximales à gauche, lorsque la séquence générée aléatoirement est extraite d'une chaîne de Markov d'ordre m ($m \geq 1$), homogène et stationnaire sur un alphabet fini.

La méthode de Chen-Stein donne uniquement une borne supérieure de la distance de variation totale entre deux lois de probabilité, elle ne nous renseigne pas sur l'ordre de grandeur de la majoration. Il peut exister d'autres majorations plus petites que la borne de Chen-Stein. Cependant, il serait intéréssant de faire des simulations afin de savoir le degré de précision de l'approximation. Ceci nécessite de trouver une majoration plus petite que la borne de Chen-Stein, et de développer des algorithmes de simulations, ce qui sera l'objet d'un nouveau travail que nous comptons faire ultérieurement, et élargir ensuite ce travail à d'autres modèles.

Bibliographie

[1] Arratia, R., Goldenstein, L., and Gordon, L. (1989). Two moments suffice for Poisson approximation : The Chen-Stein method. *Ann. Probab.*, **17** 9-25.

[2] Arratia, R., Goldenstein, L., and Gordon, L. (1990). Poisson approximation and the Chen-Stein method. *Statistical Science.* **5** 403-434.

[3] Arratia, R., Martin, D., Reinert, G. and Waterman, M. (1996). Poisson process approximation for sequence repeats and sequencing by hybridization. *J. Comp. Biol.* **3** 425-463.

[4] Barbour, A. D, , and Hall, L. (1984). On the rate Poisson convergence. *Math. Proc. Cam. Phil. Soc.* **95** 473-480.

[5] Barbour, A., Holst, L., and Janson, S. (1992). *Poisson Approximation.* Clarendon, Oxford.

[6] Chen, L.H.Y. (1975). Poisson approximation for dependent trials. *The annals of proba.* **3**, N°4, 534-545.

[7] Reinart, E., Schbath, S. (1998). Compound Poisson and Poisson approximation for occurrences of multiple words in Markov chain. *J. Comp. Biol.* **5** 223-253.

[8] Reinert, G. Schbath, S. and Waterman, M. (2005) *Applied Combinatorics on words.* volume 105 of *Encyclopedia of Mathematics and its Applications*, chapter Statistics on words with applications to biological sequences. Cambridge University Press.

[9] Roquin, E., Schbath, S. (2007). Improved compound Poisson approximation for the number of occurrences of multiple words in stationary Markov chain. *Adv. Appl. Prob.* **39** 128-140.

[10] Ruzankin, P.S. (2004). On the rate of Poisson process approximation to a Bernouilli process. *J. App. Prob.* **41**, 271-276.

[11] Schbath, S. (1995a). Compound Poisson approximation of words counts in DNA sequences. *ESAIM : Probability and Statistics.* **1** 1-16.

[12] Schbath, S. (1995b). *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V.

[13] Senoussi, R. (1990). Statistique asymptôtique presque-sûre de modèles statistiques convexes. **26** 19-44.

[14] TOUYAR, N. (2006). *Étude du nombre de répétitions maximales à gauche non chevauchantes dans des séquences d'ADN.* PhD thesis, Université de Rouen.

[15] TOUYAR, N., SCHBATH, S., CELLIER, D., DAUCHEL, H. (2008). Poisson approximation for the number of repeats in a Markov chain model. *J. Appl. Prob.* **45** 440-455.

Résumé

L'étude de la significativité statistique du nombre de répétitions chevauchantes maximales à gauche dans une séquence, est importante pour l'analyse statistique des séquences. Il est pour cela indispensable de comparer le nombre de répétitions chevauchantes maximales à gauche dans deux séquences de même longueur, l'une étant observée, l'autre aléatoire. La significativité statistique du nombre N_t^{obs} de répétitions chevauchantes maximales à gauche dans la séquence observée, requière de calculer la p-value $p = \mathbb{P}\left(N_t \geq N_t^{obs}\right)$ où N_t est le nombre de répétitions chevauchantes maximales à gauche dans la séquence aléatoire. Le but du travail développé dans cette thèse est de montrer, en utilisant la méthode de Chen-Stein, que sous certaines conditions, la loi de probabilité de N_t est approximée par une loi de Poisson de paramètre $\lambda = \mathbb{E}\left(N_t\right)$ quand la séquence aléatoire est modélisée par une chaîne de Markov homogène et stationnaire sur un alphabet fini. La détermination pratique de la p-value nécessite de connaître λ, ce qui revient à estimer λ par $\widehat{\lambda}$, nous montrons pour cela que la loi de probabilité de N_t est approximée par une loi de Poisson de paramètre $\widehat{\lambda}$. Pour une extention de la modélisation de la séquence aléatoire à une chaîne de Markov d'ordre m avec $m \geq 1$, nous montrons là aussi que l'approximation de la loi de probabilité de N_t par une loi de Poisson de paramètre $\lambda^{(m)}$ déterminé pour ce modèle, reste encore valide.

Mots clés : Nombre de répétitions chevauchantes maximales à gauche ; Approximation de Poisson ; Méthode de Chen-Stein ; Chaîne de Markov.

Abstract

The statistical significance study of the number of self-overlapping leftmost repeats in a sequence is important for the statistical analysis of sequences. It is essential for this to compare the number of self-overlapping leftmost repeats in two sequences having the same length, where one is observed while the other is random. The statistical significance of the number N_t^{obs} of self-overlapping leftmost repeats in the observed sequence requires to calculate the p-value $p = \mathbb{P}\left(N_t \geq N_t^{obs}\right)$ where N_t is the number of self-overlapping leftmost repeats in the random sequence. The aim of the work developped in this thesis is to show using the Chen-Stein method that under some conditions the probability distribution of N_t is approximated by the Poisson distribution with the parameter $\lambda = \mathbb{E}\left(N_t\right)$, when the randon sequence is generated by an homogeneous stationary Markov chain on a finite alphabet. In practice the calculus of the p-value needs to know λ, which leads to estimate λ by $\widehat{\lambda}$. We also show that the probability distribution of N_t is approximated by the Poisson distribution with the parameter $\widehat{\lambda}$. Moreover we extend the approximation to a m-order Markov chain model where $m \geq 1$, and we show that the approximation of the probability distribution of N_t by the Poisson distribution with a parameter $\lambda^{(m)}$ determined for this model, remains valid.

Keywords : Number of self-overlapping leftmost repeats ; Poisson distribution ; Chen-Stein method ; Markov chain.