UNIVERSITÉ RENÉ DESCARTES
PARIS V

THÈSE

présentée pour obtenir

Le TITRE de DOCTEUR EN SCIENCES
Discipline : Mathématique
Spécialité : Statistique

par

Sophie SCHBATH

Sujet: 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.

Soutenue le 27 octobre 1995 devant le Jury composé de :

MM.	PRUM	Bernard	Président
	BRETAGNOLLE	Jean	Rapporteur
	CHEVALET	Claude	
	EHRLICH	Dusko	
	KRICKEBERG	Klaus	
	MORGENTHALER	Stephan	Rapporteur
Mme	de TURCKHEIM	Elisabeth	

ABSTRACT

Because of many important sequencing projects, biologists have now large sets of DNA sequences from many different organisms. They need quantitative tools and statistical methods to help them in analysing sequences. Identifying words which show an important deviation between their observed frequency and their frequency predicted by a given model, is an important way to extract information from DNA sequences.

We consider different Markov chain models, either with stationary or 3-periodic transition probabilities. This last class of models is well adapted for coding sequences which are naturally split into 3-letter words.

Two different approximations for the number of occurrences of a word are used, depending on the asymptotic frame : the expected count tends to infinity or is bounded when the length of the sequence increases.

In the first part, we propose asymptotically Gaussian statistics consisting in normalised difference between the observed count and its estimated expectation, for two estimators. The main difficulty is the normalisation of this difference, with the calculation of the asymptotic variance. When the expected count is estimated by the compensator of the count related to the natural filtration of the sequence of letters, the classical central limit theorem for martingale gives the asymptotic variance. Whereas the standard deviation of the count given the sufficient statistic is a good normalisation when the expected count is estimated by the conditional expectation.

In the second part, we use the Chen-Stein method to prove the approximation of the number of occurrences for rare words by a Poisson or a compound Poisson variable. The Poisson approximation is valid for non-overlapping words, but it is not satisfactory when the word can overlap itself in the sequence. A careful study of the periodic structure of the words is needed to take into account all the overlaps.

In the last part, using these approximations, we identify words with unexpected frequency in DNA sequences of *E. coli* and *B. Subtilis*, and we propose some interpretations.

Key words : compound Poisson approximation, central limit theorem, Chen-Stein method, count of words, DNA sequences, Markov chain, periods of words, unexpected frequency.

Résumé :
Avec les multiples projets de séquençage de génomes, les biologistes ont à leur disposition une abondance de séquences d'ADN d'organismes différents. Il paraît nécessaire de leur fournir des outils divers, qu'ils soient statistiques ou informatiques, pour les aider à analyser ces séquences. La recherche de motifs dont la fréquence dans une séquence d'ADN dévie largement de celle prédite par un modèle est un problème important qui permet d'extraire de l'information de ces séquences. Pour le résoudre, il faut étudier la distribution asymptotique lorsque la longueur de la séquence tend vers l'infini, du nombre d'occurrences d'un mot dans un modèle donné. On approche alors les P-valeurs, probabilités pour que les comptages dépassent ceux observés dans la séquence.

Le modèle d'indépendance des lettres a été largement étudié mais n'est pas adapté pour rendre compte de la succession des bases dans une séquence d'ADN. Aussi, nous nous plaçons dans des modèles de chaînes de Markov d'ordre faible dont les transitions, soit sont stationnaires, soit dépendent de la position modulo 3. Ce dernier modèle est adapté lorsque la séquence est naturellement découpée en mots de longueur 3 comme les séquences codantes pour des protéines.

Suivant que l'espérance du comptage dans le modèle choisi tend vers l'infini lorsque l'on fait tendre la longueur de la séquence vers l'infini, ou au contraire reste bornée, deux approximations sont proposées.

La première partie de la thèse traite de l'approximation gaussienne du nombre d'occurrences d'un mot. Plus précisément, nous proposons des statistiques asymptotiquement gaussiennes centrées réduites construites sur la différence entre le comptage observé et une estimation du comptage dans différents modèles. Le problème délicat est la normalisation de cette différence avec notamment un calcul de variance. Deux approches sont présentées. L'une utilise des propriétés classiques pour la martingale naturelle associée au comptage et à la filtration définie par la suite des lettres de la séquence. L'estimateur du comptage est alors obtenu à partir du compensateur correspondant. Dans l'autre approche, l'estimateur est l'espérance du comptage conditionnelle à la statistique exhaustive du modèle, et l'écart type conditionnel fournit la bonne normalisation.

Dans la seconde partie, l'approximation de la loi du comptage d'un mot par une loi de Poisson ou de Poisson composée est démontrée sous l'hypothèse d'événement rare, en utilisant la méthode de Chen-Stein. Si le mot n'est pas chevauchant, l'approximation par une loi de Poisson simple est valide, par contre elle est mauvaise dès lors que le mot peut se chevaucher dans la séquence. La prise en compte de tous les chevauchements possibles nécessite une étude soigneuse de la structure périodique des mots.

Enfin, en utilisant les approximations de la loi du nombre d'occurrences d'un mot proposées dans les deux premières parties, nous analysons des séquences d'ADN des bactéries *E. coli* et *B. Subtilis*. Nous identifions des mots de fréquence exceptionnelle, tels que le motif CHI de *E. coli* ou des palindromes biologiques, et proposons quelques interprétations.

Mots clés : approximation de Poisson composée, chaîne de Markov, méthode de Chen-Stein, comptage d'un mot, fréquence inattendue, théorème de la limite centrale, période d'un mot, séquences d'ADN.

Je tiens à exprimer ici toute ma reconnaissance et mon admiration à Elisabeth de TURCKHEIM *pour la façon avec laquelle elle m'a dirigée tout au long de cette thèse. Son attention permanente, son enthousiasme et son dynamisme ont été une source immense d'encouragement et de bonne humeur. Je la remercie chaleureusement.*

Bernard PRUM *a également beaucoup contribué au bon déroulement de cette thèse en en suivant attentivement ses progressions. Aussi, je lui suis très reconnaissante de présider mon jury et lui adresse mes sincères remerciements.*

Je remercie vivement Jean BRETAGNOLLE *pour les remarques et suggestions intéressantes qu'il m'a faites en rapportant cette thèse. Sa présence dans mon jury m'honore d'autant plus qu'il fut l'un de mes professeurs.*

Stephan MORGENTHALER *a manifesté un intérêt tout particulier à mon travail. Je tiens à le remercier d'avoir accepté d'être rapporteur de cette thèse et de participer au jury.*

Je remercie également Klaus KRICKEBERG *de faire partie de mon jury.*

D'autre part, je suis particulièrement sensible à la présence de deux biologistes dans mon jury, Claude CHEVALET *et Dusko* EHRLICH*, lequel n'est pas étranger au sujet de ma thèse. J'ai eu la chance d'avoir des échanges très enrichissants avec des biologistes, ce qui fut, et sera encore, source de motivation dans mon travail.*

Je souhaite maintenant remercier tous les membres du Laboratoire de Biométrie de l'INRA à Jouy-en-Josas. En effet, durant ces trois années, j'ai eu plaisir à retrouver chaque jour cette ambiance chaleureuse et enthousiaste dont ils ont le secret.
Je remercie plus spécialement Françoise GÉLIS *qui a été d'une aide précieuse dans le développement de la boîte à outils informatique "RMES".*

Je ne peux m'empêcher d'avoir une pensée pour Anne-Sophie, Florence et Marie-Luce qui connaîtront bientôt la joie de rédiger une telle page.

Enfin, j'exprime mes plus tendres remerciements à Alex et dédie cette thèse à mes parents.

Table des matières

Chapitre 1

Présentation

Avec les multiples projets de séquençage de génomes, les biologistes ont à leur disposition une abondance de séquences d'ADN d'organismes différents. Il paraît nécessaire de leur fournir des outils divers, qu'ils soient statistiques ou informatiques, pour les aider à analyser ces séquences. Une séquence d'ADN se représente traditionnellement par la succession de ses bases Adénine, Cytosine, Guanine et Thymine. C'est pourquoi, une séquence d'ADN sera symbolisée ici par une suite de lettres $X_1X_2\cdots X_n$ où chacun des X_i est à valeurs dans l'alphabet à quatre lettres $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Nous nous sommes plus particulièrement intéressés à la recherche de motifs (suites de lettres) apparaissant dans une séquence d'ADN avec une fréquence significativement faible ou forte. L'identification de ces mots de fréquence exceptionnelle est importante pour le biologiste car, outre la signature potentielle d'un organisme, elle peut suggérer de nouvelles propriétés biologiques dont le mécanisme serait lié à la fréquence exceptionnelle de ces motifs le long du génome. On connaît par exemple des motifs rares et fréquents intervenant dans la stabilité du génome.

Le problème statistique sous-jacent est de déterminer, dans un modèle donné de la succession des lettres, la distribution, du moins asymptotique, du nombre N_W d'occurrences d'un mot W dans une séquence pour calculer ou approcher la P-valeur $\mathrm{P}\{N_W > N_W^{\mathrm{obs}}\}$. Or, le nombre N_W d'occurrences d'un mot W donné est une somme de variables aléatoires de Bernoulli Y_i qui valent 1 si une occurrence de W commence en position i, et 0 sinon. Donc intuitivement, si les variables Y_i étaient indépendantes, N_W suivrait une loi Binômiale de paramètres $(n, \mu(W))$, où n est la longueur de la séquence et $\mu(W) = \mathbb{E}Y_i$ est la probabilité d'observer un mot W en position i. Les séquences étant longues, cette loi Binômiale est approchée, soit par une loi de Poisson si $n\mu(W)$ tend vers une constante, soit par une loi Normale si $n\mu(W)$ tend vers l'infini quand n tend vers l'infini. En fait, les variables Y_i ne sont pas indépendantes dès que le mot W a plus d'une lettre. Par exemple, si un mot W de longueur h ne peut pas se chevaucher, alors s'il apparaît en position i, il ne pourra pas être observé dans les $h-1$ positions suivantes. À l'inverse, s'il peut se chevaucher c'est à dire s'il peut apparaître simultanément aux positions i

et $i+p$ avec $0 < p < h$, alors $\mathrm{P}\{Y_{i+p} = 1 \mid Y_i = 1\} > \mu(W)$. De plus, afin d'introduire plus de souplesse pour rendre compte de la succession des lettres dans la séquence, nous considérons des chaînes de Markov. Le travail consiste donc à évaluer l'influence de ces dépendances dans la loi asymptotique du comptage N_W.

La première partie de ce travail traite de l'approximation gaussienne du nombre d'occurrences N_W. Plus précisément, nous proposons des statistiques asymptotiquement gaussiennes centrées réduites, construites sur la différence entre le comptage observé d'un mot et le comptage estimé dans le modèle choisi et pour un estimateur choisi. Dans la seconde partie, nous démontrons que la loi de N_W peut être approchée par une loi de Poisson composée si l'occurrence d'un mot W est un événement rare, autrement dit si la longueur du mot W est suffisamment grande. Dans ce cas, cette approximation est plus adaptée que l'approximation gaussienne. Les résultats de ces deux premières parties sont utilisés dans la troisième partie pour identifier des mots de fréquence exceptionnelle dans des séquences d'ADN.

Choix du modèle

Le choix du modèle est important car il a une incidence sur le caractère exceptionnel d'un mot. Le modèle le plus simple consiste à choisir les lettres X_i de la séquence indépendamment les unes des autres avec probabilité $(\mu(\mathtt{A}), \mu(\mathtt{C}), \mu(\mathtt{G}), \mu(\mathtt{T}))$. Or, ce modèle largement utilisé n'est pas compatible avec la fréquence des mots de longueur 2 (Nussinov (1981), Burge *et al.* (1992)). Il est par conséquent plus raisonnable de considérer des modèles de chaîne de Markov d'ordre m qui s'ajustent sur les comptages de mots de longueur $m+1$ (modèle M_m), m restant faible toutefois pour pouvoir estimer facilement les paramètres de la chaîne (Blaisdell (1985), Brendel *et al.* (1986), Phillips *et al.* (1987), Avery (1987)). Cependant dans cette étude, le modèle reste une référence qui nous permettra d'apprécier si ce que l'on observe coïncide avec ce qui est attendu. Le but n'est pas de modéliser une séquence d'ADN mais plutôt de déterminer des écarts au modèle choisi. On verra dans la troisième partie que l'analyse d'une séquence dans les modèles successifs *M1* , *M2* , *M3*, etc ... est une bonne façon de décrire le vocabulaire d'une séquence. En effet, une partie des écarts au modèle *M1*, par exemple, vont être pris en compte dans le modèle *M2* et seront donc "orthogonaux" d'une certaine façon aux écarts dans le modèle *M2* .

Outre le modèle de chaîne de Markov d'ordre m, nous considérons deux autres modèles particulièrement adaptés aux séquences d'ADN qui codent pour des protéines. En effet, ces séquences sont naturellement découpées en codons (trois lettres successives) qui seront traduits en acides aminés via le code génétique. Aussi, il est judicieux de distinguer les bases suivant leur position par rapport aux codons, position que nous appellerons "phase". C'est pourquoi, nous avons considéré des chaînes de Markov dont les transitions de périodes 3 dépendent de la phase. On définit le modèle *Mm_3* par la donnée des transitions

$$\pi_k(a_1 a_2 \cdots a_m, a_{m+1}) = \mathrm{P}\{X_{3j+k} = a_{m+1} \mid X_{3j+k-m} = a_1, \ldots, X_{3j+k-1} = a_m\} \quad \forall j\,,$$

pour $a_1, \ldots, a_{m+1} \in \mathcal{A}$ et $k \in \{1,2,3\}$, et le modèle *M1_c* par les transitions

$$\begin{aligned}
\pi_1(a_1, a_2) &= \mathrm{P}\{X_{3j+1} = a_2 \mid X_{3j} = a_1\} \\
\pi_2(a_1, a_2) &= \mathrm{P}\{X_{3j+2} = a_2 \mid X_{3j+1} = a_1\} \\
\pi_3(a_1 a_2, a_3) &= \mathrm{P}\{X_{3j+3} = a_3 \mid X_{3j+1} = a_1,\, X_{3j+2} = a_2\}\,,
\end{aligned}$$

pour $a_1, a_2, a_3 \in \mathcal{A}$. Le modèle *Mm_3* s'ajuste ainsi sur les comptages des mots de longueur $m+1$ en phase 1, 2 et 3, tandis que le modèle *M1_c* s'ajuste sur les comptages des mots de longueur 2 en phase 1, 2 et 3 ainsi que sur les comptages des codons (mots de longueur 3 en phase 3).

Statistiques d'écart au modèle asymptotiquement gaussiennes

Une statistique naturelle pour tester si le nombre N_W d'occurrences d'un mot W est différent de celui attendu dans un modèle M donné, consiste à normaliser la différence entre N_W et $n\widehat{\mu}_M(W)$, où $\widehat{\mu}_M(W)$ est un estimateur consistant de $\mu(W)$ sous le modèle M. La difficulté réside dans le calcul de la variance de la différence, qui n'est pas la variance de N_W car $n\widehat{\mu}_M(W)$ est aussi aléatoire. Pevzner *et al.* (1989) et Stűckle *et al.* (1990) n'ont pas pris en compte cette remarque. D'autres auteurs ont approché la variance de l'écart par $n\widehat{\mu}_M(W)$ comme pour une variable aléatoire de Poisson de paramètre $n\mu(W)$ (Brendel *et al.* (1986), Avery (1987), Phillips *et al.* (1987)). Or, Prum *et al.* (1995) ont montré dans le modèle *M1* que $\widehat{\mu}_M(W)$ ne converge pas vers la variance asymptotique de $n^{-1/2}(N_W - n\widehat{\mu}_M(W))$.
Dans la première partie, nous normalisons l'écart $N_W - n\widehat{\mu}_M(W)$ en utilisant deux approches suivant l'estimateur de $n\mu(W)$ choisi dans les modèles *Mm*, *Mm_3* et *M1_c* . La première approche est basée sur le fait que la différence entre N_W et son compensateur, par rapport à la filtration naturelle engendrée par la suite des lettres de la séquence, est une martingale. On considère donc un estimateur de $n\mu(W)$ construit à partir du compensateur de N_W. Dans le cas du mot *abcd* dans *M1* par exemple, cet estimateur s'écrit

$$n\widehat{\widehat{\mu}}_1(abcd) = \frac{N_{abc}\, N_{cd}}{N_c}\ ;$$

dans le cas général, on a

$$n\widehat{\widehat{\mu}}_m(W) = \frac{N_{W^-} \times N_{\widetilde{W}}}{N_{\widetilde{W}^-}}$$

où W^- est le mot W privé de sa dernière lettre, $\widetilde{W}$ est le mot constitué des $m+1$ dernières lettres de W, et $\widetilde{W}^-$ est le mot $\widetilde{W}$ privé de sa dernière lettre. La variance asymptotique de $n^{-1/2}(N_W - n\widehat{\widehat{\mu}}_m(W))$ s'obtient alors par des théorèmes limites sur les martingales. Dans l'exemple ci-dessus, elle vaut

$$\sigma_1^2(abcd) = \frac{\mu(abcd)}{\mu(c)^2}\,(\mu(c) - \mu(cd))\,(\mu(c) - \mu(abc))\ ,$$

et la formule générale est la suivante :

$$\sigma_m^2(W) = \frac{\mu(W)}{\mu(\widetilde{W}^-)^2}\left(\mu(\widetilde{W}^-) - \mu(\widetilde{W})\right)\left(\mu(\widetilde{W}^-) - \mu(W^-)\right)\,.$$

Ces formules pour le modèle *Mm* sont un cas particulier des formules obtenues au Chapitre 3 dans le modèle *Mm_3*.
La seconde approche utilise, comme estimateur de $n\mu(W)$, l'espérance de N_W conditionnelle à la statistique exhaustive $\mathcal{S}$ du modèle, par exemple les comptages des mots de longueur $m+1$ et les m premières lettres de la séquence lorsqu'il s'agit du modèle *Mm*. L'écart type conditionnel de N_W est alors un bon facteur de normalisation. Si le mot W de longueur h s'écrit $w_1 \cdots w_h$ avec $w_i \in \mathcal{A}$, l'espérance conditionnelle de N_W dans *Mm* est asymptotiquement équivalente à l'estimateur naturel souvent utilisé par les auteurs cités plus haut

$$n\widehat{\mu}_m(W) = \frac{\prod_{j=1}^{h-m} N_{w_j \cdots w_{j+m}}}{\prod_{j=2}^{h-m} N_{w_j \cdots w_{j+m-1}}} \;;$$

pour le mot *abcd* dans *M1*, il s'écrit

$$n\widehat{\mu}_1(abcd) = \frac{N_{ab}N_{bc}N_{cd}}{N_bN_c} .$$

L'expression de la variance conditionnelle asymptotique dans *Mm*, calculée au Chapitre 2, est la suivante :

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(N_W \mid \mathcal{S}) = \mu(W) \; + 2 \sum_{d=1}^{h-m-1} \delta(W;d)\, \mu(W^{(d)}W)$$
$$+ \mu^2(W) \left(\sum_{a_1,\ldots,a_m} \frac{N^2_{a_1\cdots a_m +}(W)}{\mu(a_1\cdots a_m)} - \sum_{a_1,\ldots,a_{m+1}} \frac{N^2_{a_1\cdots a_m a_{m+1}}(W)}{\mu(a_1\cdots a_m a_{m+1})} + \frac{1-2N_{w_1\cdots w_m +}(W)}{\mu(w_1\cdots w_m)} \right) \; p.s.$$

où $\delta(W;d)$ vaut 1 si deux mots W peuvent apparaître simultanément en positions i et $i+d$ dans la séquence, et 0 sinon ; $W^{(d)}W$ est le mot composé des deux occurrences chevauchantes de W associé : $w_1 \cdots w_d w_1 \cdots w_h$. Le nombre d'occurrences de V dans le mot W est noté $N_V(W)$ et $N_{V+}(W) = \sum_{a\in\mathcal{A}} N_{Va}(W)$. On retrouve dans cette expression, le terme d'espérance $\mu(W)$ comme pour une variable de Poisson, mais aussi des termes supplémentaires non négligeables ; en effet, il y a des termes faisant apparaître les chevauchements de deux occurrences de W, et un terme en $\mu^2(W)$ qui fait intervenir l'ordre de la chaîne de Markov.

Approximation par une loi de Poisson composée pour les mots rares

Lorsque l'espérance du comptage N_W reste bornée quand on fait tendre la longueur de la séquence vers l'infini, on est dans le cadre des approximations de Poisson. En utilisant les fonctions génératrices ou la méthode de Chen-Stein, plusieurs auteurs ont établi la convergence de N_W vers une loi de Poisson dans deux cadres différents. Le cas d'un mot dans une suite de lettres indépendantes, voire identiquement distribuées, a été largement étudié (Godbole (1991), Chryssaphinou et Papastavridis (1988), Hirano et Aki (1993), Godbole et Schaffner (1993), Fu (1993)). Lorsque l'on se place sous une hypothèse markovienne, seuls des résultats pour des mots particuliers ont été présentés (Godbole (1991), Godbole et Schaffner (1993), Rajarshi (1974), Hirano et Aki (1993)). Cette restriction se retrouve également dans Geske *et al.* (1995) où une approximation par une loi de Poisson composée est étudiée.

Dès qu'un mot W peut se chevaucher, il a tendance à apparaître en "trains" d'occurrences chevauchantes dans une séquence ; l'approximation de N_W par une variable aléatoire de Poisson simple est alors mauvaise. Par construction, les trains ne se chevauchent pas dans la séquence. Ainsi, nous démontrons dans la seconde partie que le nombre de trains peut être approché par une variable aléatoire de Poisson, et que l'on peut approcher la loi de N_W par une loi de Poisson composée, en utilisant la méthode de Chen-Stein. Les composantes de cette loi de Poisson composée sont directement reliées aux trains composés d'exactement k occurrences chevauchantes de W. Par conséquent, l'étude de mots W quelconques nécessite une description de la structure complexe des trains de W, qui n'est possible que si l'on connaît la structure périodique des mots.
Ainsi, ce travail passe par l'étude des périodes des mots. Un entier p est une période du mot $W = w_1 w_2 \cdots w_h$ si $w_i = w_{i+p}$ pour $i = 1 \cdots h - p$. En utilisant l'ensemble $\mathcal{P}'(W)$ des périodes de W qui ne sont pas des multiples de la période minimale de W, on peut calculer l'espérance de l'occurrence d'un train de W, donnée par

$$\widetilde{\mu}(W) = \mu(W) \; - \sum_{p \in \mathcal{P}'(W)} \mu(W^{(p)} W) \; ;$$

pour le mot `AACAACAA`, par exemple, qui possède les trois périodes 3, 6 et 7, cette expression s'écrit

$$\widetilde{\mu}(\mathtt{AACAACAA}) = \mu(\mathbf{A}\mathtt{ACAACAA}) - \mu(\mathbf{A}\mathtt{AC}\mathbf{A}\mathtt{ACAACAA}) - \mu(\mathbf{A}\mathtt{ACAACA}\mathbf{A}\mathtt{ACAACAA}) \, .$$

On peut aussi calculer l'espérance de l'occurrence d'un train d'exactement k occurrences de W :

$$\widetilde{\mu}_k(W) = \sum_{C \in \mathcal{C}_k} \mu(C) \; - \; 2 \sum_{C' \in \mathcal{C}_{k+1}} \mu(C') \; + \sum_{C'' \in \mathcal{C}_{k+2}} \mu(C'') \, ,$$

où $\mathcal{C}_k$ désigne l'ensemble de tous les motifs composés d'exactement k occurrences chevauchantes de W. Les paramètres de la loi de Poisson composée sont précisément les quantités
$(n-h+1)\widetilde{\mu}_k(W)$. Le calcul de $\widetilde{\mu}_1(\mathtt{AACAA})$, par exemple, fait intervenir les mots composés suivants : $\mathcal{C}_1 = \{$**A**ACAA$\}$, $\mathcal{C}_2 = \{$**A**AC**A**ACAA, **A**ACA**A**ACAA$\}$ et $\mathcal{C}_3 = \{$**A**AC**A**AC**A**ACAA, **A**AC**A**ACA**A**ACAA, **A**ACA**A**AC**A**ACAA, **A**ACA**A**ACA**A**ACAA$\}$; les lettres en gras indiquent le début des occurrences chevauchantes de $W =$ `AACAA`.

La loi de Poisson composée limite, mise en évidence au Chapitre 5, se réduit à une loi de Poisson uniquement pour les mots non chevauchants. On peut constater, sur la Figure 1.1, la différence entre la loi de Poisson composée approchant la loi de N_W et la loi de Poisson de paramètre $n\widehat{\mu}_M(W)$, pour le cas extrême des mots chevauchants $W =$`AAAAAA` et $W =$`AAAAAAAA`.

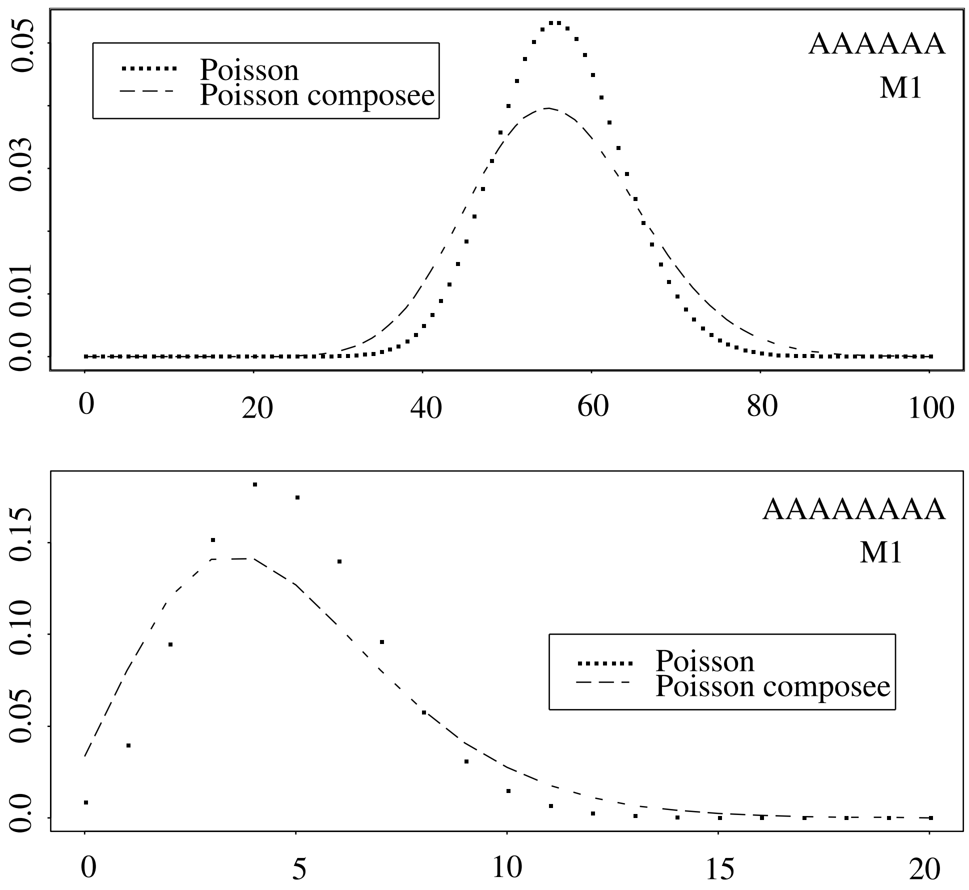


FIG. 1.1 – *Lois d'approximations de Poisson et de Poisson composée, sous le modèle M1, pour les mots* AAAAAA *et* AAAAAAAA *dans une séquence de 111402 bases (E. coli).*

Il est donc indispensable de faire une approximation de Poisson composée plutôt qu'une approximation par une loi de Poisson.

Applications et conclusion

Dans la troisième partie, nous analysons des séquences d'ADN de deux bactéries *E. coli* et *B. subtilis*, ainsi que deux bactériophages *Lambda* et *T7*, parasites de *E. coli*. Pour une séquence donnée, un modèle choisi et une longueur de mots fixée, nous calculons les P-valeurs $\mathrm{P}\{N_W > N_W^{\mathrm{obs}}\}$.
Pour des séquences de l'ordre de 100000 bases et des mots de longueur inférieure à huit, les P-valeurs calculées par l'approximation gaussienne sont légèrement différentes de celles obtenues par la loi de Poisson composée, mais les mots exceptionnels restent les mêmes. Un calcul plus fin des P-valeurs par les techniques de grandes déviations, pour les mots exceptionnels, semble valider l'utilisation de la loi de Poisson composée, mais celle-ci est plus difficile à mettre en oeuvre.

Une analyse de séquences codantes de *E. coli* et *B. subtilis* montre que, dans le modèle $M1\_3$ qui s'ajuste sur les comptages des mots de longueur 2 en phase 1, 2 et 3, les mots de longueur 3 exceptionnels sont essentiellement les codons, c'est à dire ceux en phase 3. Le modèle $M1\_c$ est donc parcimonieux entre $M1\_3$ et $M2\_3$, pour extraire de ces séquences de l'information "orthogonale" au biais de codons, bien connu des biologistes.
Chez *E. coli*, on montre que les palindromes[1] biologiques sont exceptionnellement rares, surtout ceux de longueur 6. Une explication peut être donnée en rapport avec les sites de restriction qui sont des motifs, souvent palindromiques, reconnus par des enzymes de restriction des bactéries et aux occurrences desquels ces enzymes coupent le génome. Il est donc naturel que de tels sites soient évités dans le génome de *E. coli*.
À l'inverse, le motif CHI de *E. coli*, `GCTGGTGG`, dont la présence est bénéfique le long du génome de *E. coli*, est exceptionnellement fréquent dans tous les modèles de $M0$ à $M6$. Aucun motif ne possède cette propriété statistique chez *B. subtilis* dont la séquence CHI reste encore inconnue, tandis que deux mots de longueur 8 présentent cette caractéristique chez le phage *T7*.

Ces analyses ont été réalisées avec la boîte à outils informatique "RMES" que nous avons développée en vue de la mettre à disposition des biologistes.

Ce travail statistique permet donc de mettre en évidence des motifs dont la fréquence dévie largement de celle prédite par un modèle choisi. Le rôle éventuel, dans le génome d'un organisme, de ces motifs suggérés par une analyse statistique ne peut être identifié qu'au moyen d'expérimentations.

1. Un palindrome biologique est un mot sur l'alphabet {`A`, `C`, `G`, `T`} qui se lit de la même façon de gauche à droite, et de droite à gauche en complémentant les bases `A` par `T`, `T` par `A`, `C` par `G` et `G` par `C` ; par exemple, `AATT` et `CATG` sont deux palindromes de longueur 4.

Première partie

Approximation gaussienne du comptage d'un mot dans une séquence d'ADN

Chapitre 2

Modèle markovien d'ordre m

2.1 Introduction

Le travail présenté dans ce chapitre généralise, au cas d'une chaîne de Markov d'ordre m quelconque (modèle Mm), la méthode conditionnelle proposée par Prum *et al.* (1995) pour $m = 1$. Cette méthode consiste à recentrer le comptage N_W d'un mot W par son espérance conditionnelle à la statistique exhaustive du modèle Mm, puis à normaliser cet écart par la racine carrée de la variance conditionnelle. Nous calculons celle-ci, tout comme l'espérance conditionnelle, à l'aide de la formule de Whittle (1955), qui donne le nombre de séquences ayant leur première lettre et des comptages de mots de longueur 2 fixés, et d'une idée de Cowan (1991). Ce dernier établit une bijection entre des séquences contenant un mot W en position i et celles obtenues en supprimant l'intérieur de ces mots W en i.
L'extension de ces techniques au cas $m > 1$ nécessite un certain nombre de développements qui font l'objet de ce chapitre. En effet, il y a une différence notable entre le cas $m = 1$ et $m > 1$. Nous présentons directement le cas $m = 2$ en reprenant la méthode utilisée dans Prum *et al.* (1995) mais dans le cadre différent nécessaire au cas $m = 2$. Comme nous le verrons, la méthode dans le cas $m = 2$ s'étend sans difficulté au cas général et permet d'avoir des notations simples pour présenter la méthode générale.

2.2 Présentation de la méthode conditionnelle dans le modèle M2

2.2.1 Cadre

Une séquence d'ADN de longueur n se représente naturellement par la succession de ses n bases, chacune de ces bases pouvant être `A`, `C`, `G` ou `T`. On considère donc ici, une séquence $S = X_1X_2 \cdots X_n$ générée par une chaîne de Markov stationnaire d'ordre 2, sur l'alphabet $\mathcal{A} =$

$\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$. Sa matrice de transition $\Pi = (\pi(ab,c))_{a,b,c\in\mathcal{A}}$ est définie par

$$\pi(ab,c) = \mathrm{P}\{X_{i+2} = c \mid X_{i+1} = b, X_i = a\}.$$

On suppose que $\pi(ab,c) > 0$ pour tout $a,b,c \in \mathcal{A}$, pour assurer l'ergodicité de la chaîne de Markov ; μ désignera ainsi l'unique mesure invariante.

Le nombre N_W d'occurrences d'un mot $W = w_1w_2\cdots w_h$ de longueur h, dans la séquence S, s'écrit

$$N_W = N(W) = N(w_1w_2\cdots w_h) = \sum_{i=1}^{n-h+1} \mathbb{1}\{X_i = w_1,\ldots,X_{i+h-1} = w_h\},$$

où $\mathbb{1}\{\cdot\}$ est la fonction indicatrice. On note $\mu(W) = \mathrm{P}\{X_i = w_1,\ldots,X_{i+h-1} = w_h\}$.

On se propose d'étudier le comportement asymptotique de la statistique centrée réduite

$$U_2(W) = \frac{N_W - \mathbb{E}(N_W \mid \mathcal{S}_2)}{\mathrm{Var}(N_W \mid \mathcal{S}_2)^{1/2}}$$

où $\mathcal{S}_2$ est la statistique exhaustive du modèle, c'est à dire

$$\mathcal{S}_2 = \{X_1X_2,\ (N_{abc}), a,b,c \in \mathcal{A}\}\ ;$$

X_1 et X_2 sont les deux premières lettres de S et N_{abc} est le nombre d'occurrences du 3-mot abc dans S. On considérera plutôt $\mathcal{S}_2$ comme l'ensemble de toutes les séquences de longueur n sur l'alphabet $\mathcal{A}$ ayant leurs deux premières lettres identiques à celles de S, X_1 et X_2, ainsi que les mêmes comptages de mots abc que dans S, c'est à dire N_{abc}.

Pour cela, on se ramène au cas d'une chaîne de Markov d'ordre 1. En effet, la séquence $S' = X'_1X'_2\cdots X'_{n-1}$, où $X'_i = X_iX_{i+1}$, est une chaîne de Markov d'ordre 1 sur l'espace d'état fini $\mathcal{A}^2$. Sa matrice de transition $\Pi' = (\pi'(A,B))_{A,B\in\mathcal{A}^2}$ est définie par

$$\pi'(A,B) = \begin{cases} \pi(ab,c) & \text{si } A = ab \text{ et } B = bc,\\ 0 & \text{sinon.}\end{cases}$$

Par convention, on notera les "lettres" de l'alphabet $\mathcal{A}^2$ en majuscules.
Bien que cette chaîne de Markov possède des transitions nulles, elle a les bonnes propriétés suivantes :

- elle est irréductible : $\forall A,B \in \mathcal{A}^2$, $\exists \ell$ tel que $\Pi'^{\ell}(A,B) > 0$.
  En effet, si $A = ab$ et $B = cd$ alors $\Pi'^2(A,B) = \pi(ab,c)\pi(bc,d) > 0$.

- Elle est apériodique : $\mathrm{pgcd}\{\ell \mid \Pi'^{\ell}(A,B) > 0\} = 1$.
  En effet, $\Pi'^2(A,B) > 0$ et, si $A = ab$ et $B = cd$, $\Pi'^3(A,B) = \pi(ab,c)\pi(bc,c)\pi(cc,d) > 0$.

Par conséquent, S' est récurrente positive et ergodique sur $\mathcal{A}^2$. La mesure invariante μ' sur $\mathcal{A}^2$ est telle que $\mu'(A) = \mu(ab)$ si $A = ab$; par abus de notation on l'écrira $\mu(A)$. Les théorèmes limites classiques, du type loi des grands nombres et théorème de la limite centrale, s'appliquent donc (Dacunha-Castelle et Duflo (1983), paragraphe 4.4.1).

On dit qu'un mot de longueur 2 (2-mot) AB sur l'alphabet $\mathcal{A}^2$ est *compatible* si et seulement si il existe des lettres a, b, c dans $\mathcal{A}$ telles que $A = ab$ et $B = bc$.

Le nombre $N_{AB}(S')$ d'occurrences du 2-mot AB dans la séquence S', s'obtient de la façon suivante :

$$N_{AB}(S') = \begin{cases} N_{abc} & \text{si } A = ab \text{ et } B = bc, \text{ (}AB\text{ compatible)} \\ 0 & \text{sinon (}AB\text{ non compatible).} \end{cases}$$

On peut ainsi associer à la séquence S', l'ensemble $\mathbb{S}$ de toutes les séquences de longueur $n-1$ sur l'alphabet $\mathcal{A}^2$ ayant leur première lettre égale à celle de S' (X'_1) et des comptages de 2-mots AB égaux aux $N_{AB}(S')$. On désignera par $N_{AB}(\mathbb{S})$, le nombre de 2-mots AB dans les séquences de $\mathbb{S}$, autrement dit $N_{AB}(\mathbb{S}) = N_{AB}(S')$, que l'on notera sans ambiguïté N_{AB}.

Enfin, au mot $W = w_1 w_2 \cdots w_h$ écrit dans l'alphabet $\mathcal{A}$, correspond le mot $\mathbb{W} = W_1 W_2 \cdots W_{h-1}$ écrit dans l'alphabet $\mathcal{A}^2$, avec $W_i = w_i w_{i+1}$. Ainsi, par construction de la séquence S', le nombre d'occurrences du mot $\mathbb{W}$ dans S', noté sans ambiguïté $N_{\mathbb{W}}$, est égal au nombre N_W d'occurrences du mot W dans S. Par conséquent, on a les égalités suivantes :

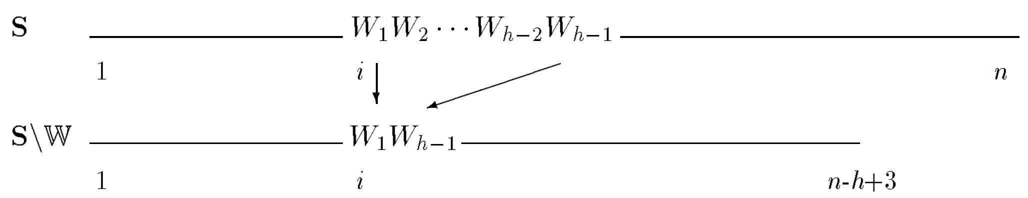
$$\mathbb{E}(N_W \mid \mathcal{S}_2) = \mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}), \tag{2.1}$$

$$\mathrm{Var}(N_W \mid \mathcal{S}_2) = \mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}), \tag{2.2}$$

dont les deux membres de droite se calculent respectivement à l'aide des formules de Cowan (1991) et Whittle (1955) que nous reprenons ci-dessous.

2.2.2 Espérance et variance conditionnelles du comptage N_W

Nous rappelons ici brièvement comment s'obtiennent les expressions de $\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S})$ et $\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S})$ en utilisant une idée de Cowan (1991) et la formule de Whittle (1955) qui donne le cardinal de l'ensemble $\mathbb{S}$, noté $\#\mathbb{S}$. L'idée est la suivante : on établit une bijection entre les séquences $\mathbf{S}$ de $\mathbb{S}$ qui ont un mot $\mathbb{W}$ en position i, et celles, notées $\mathbf{S}\backslash\mathbb{W}$, obtenues à partir des précédentes en supprimant l'intérieur du mot $\mathbb{W}$ en i ; comme le montre le schéma ci-dessous, ces dernières possèdent un mot $W_1 W_{h-1}$ en position i et sont de longueur $n-h+3$.



On définit ainsi l'ensemble $\mathbb{S}\backslash\mathbb{W}$ des séquences de longueur $n-h+3$ ayant des comptages de AB corrigés, pour prendre en compte la délétion des lettres $W_2 W_3 \cdots W_{h-2}$ d'un mot $\mathbb{W}$, de la façon suivante :

$$N_{AB}(\mathbb{S}\backslash\mathbb{W}) = N_{AB} - n_{AB} + \mathbb{1}\{AB = W_1 W_{h-1}\}\,,$$

où n_{AB} est le nombre de AB dans $\mathbb{W}$. Par simplicité, on notera d_{AB} la correction $n_{AB} - \mathbb{1}\{AB = W_1 W_{h-1}\}$, et on pose $U = W_1$ et $V = W_{h-1}$.

Remarque : Par construction, le mot $\mathbb{W}$ et toutes les séquences de $\mathbb{S}$ ne contiennent que des 2-mots compatibles. Par contre, les séquences de $\mathbb{S}\backslash\mathbb{W}$ peuvent éventuellement posséder un 2-mot non compatible, celui-ci étant unique. En effet, il se peut que le mot UV créé par délétion ne soit pas compatible, auquel cas $N_{UV}(\mathbb{S}\backslash\mathbb{W}) = 1$. Dans ce cas, notre cadre asymptotique est différent de celui de Prum *et al.* (1995) pour lequel les comptages des mots de longueur 2 tendent vers l'infini lorsque l'on fait tendre la longueur de la séquence vers l'infini. Ainsi, dans la suite, on traitera toujours séparément le cas où UV est compatible, du cas où UV ne l'est pas.

Le calcul de l'espérance conditionnelle s'effectue alors de la manière suivante

$$\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) = \sum_i \mathrm{P}\{S' \text{ a un mot } \mathbb{W} \text{ en } i \mid \mathbb{S}\}.$$

En remarquant que toutes les séquences de $\mathbb{S}$ (resp. $\mathbb{S}\backslash\mathbb{W}$) sont équiprobables de probabilité $(\#\mathbb{S})^{-1}$ (resp. $(\#(\mathbb{S}\backslash\mathbb{W}))^{-1}$), on a

$$\mathrm{P}\{S' \text{ a un mot } \mathbb{W} \text{ en } i \mid \mathbb{S}\} = \frac{\#\{\text{séquences de } \mathbb{S} \text{ ayant un mot } \mathbb{W} \text{ en } i\}}{\#\mathbb{S}}\,,$$

d'où

$$\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) = \frac{1}{\#\mathbb{S}} \sum_i \#\{\text{séquences de } \mathbb{S} \text{ ayant un mot } \mathbb{W} \text{ en } i\}\,.$$

De plus, la bijection implique qu'il y a autant de séquences dans $\mathbb{S}$ qui possèdent un mot $\mathbb{W}$ en i, que de séquences dans $\mathbb{S}\backslash\mathbb{W}$ qui ont un mot UV en i. Par conséquent,

$$\#\mathbb{S} \times \mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) = \#(\mathbb{S}\backslash\mathbb{W}) \times \mathbb{E}(N_{UV} \mid \mathbb{S}\backslash\mathbb{W}).$$

Or, $\mathbb{E}(N_{UV} \mid \mathbb{S}\backslash\mathbb{W})$ n'est pas aléatoire et vaut précisément $N_{UV}(\mathbb{S}\backslash\mathbb{W})$, c'est à dire $N_{UV} - n_{UV} + 1$; on obtient donc

$$\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) = \frac{\#(\mathbb{S}\backslash\mathbb{W})}{\#\mathbb{S}} \left(N_{UV} - n_{UV} + 1\right)\,. \tag{2.3}$$

L'expression de la variance conditionnelle s'obtient de la même façon, en prenant toutefois en compte la possibilité qu'a un mot de se chevaucher dans la séquence. Une bijection sera établie ici entre l'ensemble des séquences de $\mathbb{S}$ qui ont deux mots $\mathbb{W}$ aux positions i et j avec $|i-j| \geq h-2$ (c'est à dire qu'ils se chevauchent sur une lettre au plus) et l'ensemble des séquences de $\mathbb{S}\backslash\mathbb{W}\backslash\mathbb{W}$ qui ont deux mots UV aux positions i et $j-h+3$. L'ensemble $\mathbb{S}\backslash\mathbb{W}\backslash\mathbb{W}$ s'obtient cette fois en

corrigeant les N_{AB} par $2d_{AB}$ pour tenir compte de la délétion des lettres $W_2W_3\cdots W_{h-2}$ de deux mots $\mathbb{W}$ se chevauchant sur une lettre au plus.
La formule donnée par Prum *et al.* (1995) est la suivante

$$\begin{aligned}\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}) &= \mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) \\ &+ N_{UV}(\mathbb{S}\setminus\mathbb{W}\setminus\mathbb{W})(N_{UV}(\mathbb{S}\setminus\mathbb{W}\setminus\mathbb{W})-1)\,\frac{\#(\mathbb{S}\setminus\mathbb{W}\setminus\mathbb{W})}{\#\mathbb{S}} \\ &+ 2\sum_{d=1}^{h-3}\delta(\mathbb{W};d)N_{UV}(\mathbb{S}\setminus\mathbb{W}^{(d)}\mathbb{W})\,\frac{\#(\mathbb{S}\setminus\mathbb{W}^{(d)}\mathbb{W})}{\#\mathbb{S}} \\ &- \mathbb{E}(N_{\mathbb{W}}\mid\mathbb{S})^2\,, \end{aligned} \tag{2.4}$$

où $\delta(\mathbb{W};d)=1$ si les $h-1-d$ premières lettres de $\mathbb{W}$ sont identiques aux $h-1-d$ dernières et on note alors $\mathbb{W}^{(d)}\mathbb{W}$ le mot chevauché $W_1W_2\cdots W_dW_1W_2\cdots W_{h-1}$; sinon $\delta(\mathbb{W};d)=0$. Cet indicateur $\delta(\mathbb{W};d)$ gère ainsi les chevauchements de $\mathbb{W}$ avec lui-même (d correspond au décalage entre les deux occurrences chevauchantes). Une description exhaustive de ces chevauchements sera faite dans la Partie II.
Un calcul analogue de variance conditionnelle sera davantage détaillé au Chapitre 3.

2.3 Résultats asymptotiques dans le modèle M2

Nous énonçons ci-dessous le résultat essentiel de ce chapitre, que nous démontrons en utilisant plusieurs lemmes. Lorsque UV est compatible, nous reprenons les techniques de démonstration de Prum *et al.* (1995), mais celles-ci ne s'appliquent pas si UV n'est pas compatible.

Théorème 1 *La statistique centrée réduite*

$$U_2(W)=\frac{N_W-\mathbb{E}(N_W\mid\mathcal{S}_2)}{Var(N_W\mid\mathcal{S}_2)^{1/2}}$$

est définie pour n assez grand et converge en loi vers une gaussienne centrée réduite.

Preuve : La première étape consiste à établir la convergence en loi de $n^{-1/2}(N_W-n\widehat{\mu}_2(W))$ vers une gaussienne centrée, où $n\widehat{\mu}_2(W)$ est l'estimateur naturel de $\mathbb{E}N_W$ suivant :

$$n\widehat{\mu}_2(W)=\frac{\prod_{j=1}^{h-2}N(w_jw_{j+1}w_{j+2})}{\prod_{j=2}^{h-2}N(w_jw_{j+1})}\,.$$

Pour cela, on écrit $n^{-1/2}(N_W-n\widehat{\mu}_2(W)) = n^{-1/2}f(\widetilde{N})$ avec $\widetilde{N}=(N_W,N(w_1w_2w_3),\ldots,N(w_{h-2}w_{h-1}w_h),N(w_2w_3),\ldots,N(w_{h-2}w_{h-1}))$, et on fait une approximation de premier ordre de $f(\widetilde{N})$ au voisinage de $f(\mathbb{E}\widetilde{N})$, car f est dérivable pour n suffisamment grand ; en effet, à partir d'un certain rang, tous les comptages $N(ab)$ sont strictement positifs et la fonction f est bien définie. En remarquant que $f(\mathbb{E}\widetilde{N})=0$ et que chaque composante $n^{-1/2}(\widetilde{N}_j-\mathbb{E}\widetilde{N}_j)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\sigma_j^2)$

d'après un théorème de limite centrale pour les chaînes de Markov (Dacunha-Castelle et Duflo (1983), Théorème 4.3.18), on obtient la normalité asymptotique de $n^{-1/2}(N_W - n\widehat{\mu}_2(W))$ annoncée. On montre ensuite (Lemme 2) que

$$\mathbb{E}(N_W \mid \mathcal{S}_2) = n\widehat{\mu}_2(W) + \mathrm{O}(1) \quad p.s.$$

ce qui nous assure la convergence en loi de $n^{-1/2}(N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))$ vers une gaussienne centrée. Pour montrer que la limite de $n^{-1}\mathrm{Var}(N_W \mid \mathcal{S}_2)$ est bien la variance limite, il suffit de vérifier d'une part, que $n^{-1}\mathrm{Var}(N_W \mid \mathcal{S}_2)$ a une limite presque sûre (Lemme 3), et d'autre part que $n^{-1/2}(N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))$ possède un moment d'ordre strictement supérieur à 2 borné. Pour cela, on s'assure que le moment conditionnel d'ordre 4 de $n^{-1/2}(N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))$ est presque sûrement borné (Lemme 5 en Annexe A).
On obtient ainsi la convergence en loi de U_W vers la loi $\mathcal{N}(0,1)$.

Le fait que la statistique U_W soit bien définie à partir d'un certain rang vient de la stricte positivité de la limite de $n^{-1}\mathrm{Var}(N_W \mid \mathcal{S}_2)$ donnée au Lemme 3. On vérifie analytiquement que cette limite est strictement positive pour tous les mots de longueur 4 (voir la formule (2.8)) à partir du moment où aucune probabilité de transition n'est égale à 0, ce qui est notre hypothèse. Ce résultat se généralise en fait au cas d'un mot W de longueur h dans le modèle $Mh\text{-}2$ d'ordre $h-2$. Par conséquent, si W est de longueur h supérieure à 4, il suffit de remarquer que
$\mathrm{Var}(N_W \mid \mathcal{S}_2) \geq \mathrm{Var}(N_W \mid \mathcal{S}_{h-2})$ car $\mathcal{S}_2 \supset \mathcal{S}_{h-2}$, d'où

$$\lim n^{-1}\mathrm{Var}(N_W \mid \mathcal{S}_2) \geq \lim n^{-1}\mathrm{Var}(N_W \mid \mathcal{S}_{h-2}) > 0\,.$$

□

Remarque : Pour s'assurer que le moment d'ordre 4 de $n^{-1/2}(N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))$ est borné, on peut faire l'économie du calcul du moment conditionnel d'ordre 4 ; ceci est développé en Annexe A.4.

Avant de calculer l'espérance et la variance conditionnelles asymptotiques (Lemmes 2 et 3), on est amené à étudier le comportement asymptotique des rapports de cardinaux $[\#(\mathbb{S}\backslash\mathbb{W}) \,/\, \#\mathbb{S}]$ et $[\#(\mathbb{S}\backslash\mathbb{W}\backslash\mathbb{W}) \,/\, \#\mathbb{S}]$ qui apparaissent dans les expressions (2.3) et (2.4) de l'espérance et la variance conditionnelles.

Comportement asymptotique des rapports de cardinaux

On rappelle la formule de Whittle (1955) qui donne l'expression du cardinal de $\mathbb{S}$ en fonction des comptages N_{AB}.

Théorème 2 (Whittle, 1955)

$$\#\mathbb{S} = K(\mathbb{S})\, H_{X'_{n-1}}(\mathbb{S})$$

avec

$$K(\mathbb{S}) = \prod_A \frac{N_{A+}\,!}{\prod_B N_{AB}\,!}$$

où $N_{A+} = \sum_B N_{AB}$, H_r *est le cofacteur de n'importe quel élément de la ligne* r *de la matrice* $I - \hat{\Pi}$ *et* $\hat{\Pi} = (\hat{\pi}(A,B))_{A,B\in\mathcal{A}^2}$ *est la matrice de transition empirique, c'est à dire* $\hat{\pi}(A,B) = N_{AB}/N_{A+}$.

On notera de la même façon $d_{A+} = \sum_B d_{AB}$.

$$\text{Soient} \qquad R_1 = \frac{\#(\mathbb{S}\backslash\mathbb{W})}{\#\mathbb{S}}, \qquad R_2 = \frac{\#(\mathbb{S}\backslash\mathbb{W}\backslash\mathbb{W})}{\#\mathbb{S}}, \qquad R_3 = \frac{\#(\mathbb{S}\backslash\mathbb{W}\backslash\mathbb{W}\backslash\mathbb{W})}{\#\mathbb{S}}, \ldots$$

Le lemme ci-dessous donne le comportement asymptotique de R_k lorsque n tend vers l'infini ; en effet, si AB est compatible alors $\pi(A,B) > 0$ et N_{AB} tend vers l'infini avec n. De plus, on a les convergences presque sûres suivantes

$$\frac{N_{AB}}{N_{A+}} \xrightarrow{ps} \pi(A,B) \quad \text{et} \quad \frac{N_{A+}}{n} \xrightarrow{ps} \mu(A).$$

Lorsque UV n'est pas compatible, le résultat du lemme suivant diffère de celui démontré par Prum *et al.* (1995).

Lemme 1 : Comportement asymptotique de R_k. *Lorsque* n *tend vers l'infini,*

$$R_k = \left(\frac{(N_{U+})^k}{k!}\right)^{1-\delta} C^k \left(1 + \frac{\Delta_k}{n} + o(\frac{1}{n})\right) \quad p.s.$$

avec

$$\delta = 1\!\!1\{UV \text{ est compatible }\}$$

$$C = \prod_{\substack{AB\\ compatibles}} \left(\frac{N_{AB}}{N_{A+}}\right)^{d_{AB}} \xrightarrow{ps} \delta\,\frac{\mu(\mathbb{W})}{\mu(UV)} + (1-\delta)\frac{\mu(\mathbb{W})}{\mu(U)}$$

et

$$\Delta_k = \Delta_k^* + \Delta_k^{**}$$

où Δ_k^* *est obtenu lors du développement asymptotique de* $K(\mathbb{S}\backslash\mathbb{W}\cdots\backslash\mathbb{W})/K(\mathbb{S})$,

$$\Delta_k^* = \sum_A \frac{kd_{A+}(kd_{A+}-1)}{2}\frac{n}{N_{A+}} - \sum_{\substack{AB\\ compatibles}} \frac{kd_{AB}(kd_{AB}-1)}{2}\frac{n}{N_{AB}},$$

et Δ_k^{**} *est obtenu lors du développement asymptotique de* $H(\mathbb{S}\backslash\mathbb{W}\cdots\backslash\mathbb{W})/H(\mathbb{S})$ *et satisfait*

$$\Delta_k^{**} = k\Delta_1^{**}.$$

Preuve : Le comportement asymptotique de R_k est démontré en Annexe A.1. Nous calculons ici la limite presque sûre de C, en distinguant deux cas.

- Si UV est compatible ($\delta = 1$),

$$C = \prod_{AB\; comp} \left(\frac{N_{AB}}{N_{A+}}\right)^{n_{AB}-\mathbb{1}\{AB=UV\}} = \frac{N_{U+}}{N_{UV}} \prod_{AB\; comp} \left(\frac{N_{AB}}{N_{A+}}\right)^{n_{AB}}$$

  Par conséquent,

$$C \xrightarrow{ps} \frac{1}{\pi(U,V)} \prod_{AB\; comp} \pi(A,B)^{n_{AB}} = \frac{1}{\pi(U,V)} \cdot \frac{\mu(\mathbb{W})}{\mu(U)},$$

  et on utilise que $\mu(U)\pi(U,V) = \mu(UV)$.

- Si UV n'est pas compatible ($\delta = 0$),

$$C = \prod_{AB\; comp} \left(\frac{N_{AB}}{N_{A+}}\right)^{n_{AB}} \xrightarrow{ps} \frac{\mu(\mathbb{W})}{\mu(U)}$$

□

2.3.1 Espérance conditionnelle asymptotique de N_W

Le lemme suivant montre que l'espérance conditionnelle de N_W est, presque sûrement, asymptotiquement équivalente à l'estimateur naturel $n\widehat{\mu}_2(W)$ de l'espérance de N_W.

Lemme 2 : Espérance conditionnelle asymptotique dans $M2$

$$\mathbb{E}(N_W \mid \mathcal{S}_2) = n\widehat{\mu}_2(W) + \ O(1) \quad p.s.$$

Preuve : on utilise le Lemme 1 pour faire un développement asymptotique de l'espérance conditionnelle donnée par la formule (2.3) :

$$\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) = \frac{\#(\mathbb{S}\backslash\mathbb{W})}{\#\mathbb{S}}(N_{UV} - n_{UV} + 1)\,.$$

On distingue deux cas :

- UV est compatible ($\delta = 1$),

$$\begin{aligned}
\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) &= N_{UV} \prod_{AB\; comp} \left(\frac{N_{AB}}{N_{A+}}\right)^{d_{AB}} + \ \mathrm{O}(1) \\
&= N_{UV} \cdot \frac{N_{U+}}{N_{UV}} \cdot \frac{\prod_{j=1}^{h-2} N(W_j W_{j+1})}{\prod_{j=1}^{h-2} N(W_j+)} + \ \mathrm{O}(1) \\
&= \frac{\prod_{j=1}^{h-2} N(W_j W_{j+1})}{\prod_{j=2}^{h-2} N(W_j)} + \ \mathrm{O}(1)
\end{aligned}$$

  car $N(W_j) = N(W_j+) + \mathbb{1}\{X'_{n-1} = W_j\}$.

- UV n'est pas compatible ($\delta = 0$),

$$\begin{aligned}
\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) &= N_{U+} \prod_{AB\ comp} \left(\frac{N_{AB}}{N_{A+}}\right)^{d_{AB}} + \mathrm{O}(1) \\
&= N_{U+} \cdot \frac{\prod_{j=1}^{h-2} N(W_j W_{j+1})}{\prod_{j=1}^{h-2} N(W_j +)} + \mathrm{O}(1) \\
&= \frac{\prod_{j=1}^{h-2} N(W_j W_{j+1})}{\prod_{j=2}^{h-2} N(W_j)} + \mathrm{O}(1)\,.
\end{aligned}$$

Par conséquent, dans les deux cas

$$\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) = \frac{\prod_{j=1}^{h-2} N(W_j W_{j+1})}{\prod_{j=2}^{h-2} N(W_j)} + \mathrm{O}(1) \quad p.s.$$

Or

$$\frac{\prod_{j=1}^{h-2} N(W_j W_{j+1})}{\prod_{j=2}^{h-2} N(W_j)} = \frac{\prod_{j=1}^{h-2} N(w_j w_{j+1} w_{j+2})}{\prod_{j=2}^{h-2} N(w_j w_{j+1})} = n\widehat{\mu}_2(W)$$

et $\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) = \mathbb{E}(N_W \mid \mathcal{S}_2)$, ce qui démontre le lemme. □

2.3.2 Variance conditionnelle asymptotique de N_W

Pour calculer la limite presque sûre de $\frac{1}{n}\mathrm{Var}(N_W \mid \mathcal{S}_2)$ lorsque n tend vers l'infini, nous commençons par calculer celle de $\frac{1}{n}\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S})$, puis nous transformons celle-ci pour l'écrire avec l'alphabet $\mathcal{A}$. En reprenant la formule (2.4) de $\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S})$, nous utilisons le Lemme 1 en distinguant les deux cas :

- si UV est compatible,

$$\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}) - 2\sum_{d=1}^{h-3} \delta(\mathbb{W}; d)\, \mathbb{E}(N_{\mathbb{W}(d)\mathbb{W}} \mid \mathbb{S}) =$$

$$N_{UV}\, C + \frac{N_{UV}^2}{n}\, C^2(\Delta_2 - 2\Delta_1) - N_{UV}\, C^2(2d_{UV} + 1) + \mathrm{O}(1)\,,$$

or

$$\begin{aligned}
\Delta_2 - 2\Delta_1 &= \Delta_2^* - 2\Delta_1^* \\
&= n\left(\sum_A \frac{d_{A+}^2}{N_{A+}} - \sum_{AB\ comp} \frac{d_{AB}^2}{N_{AB}}\right).
\end{aligned}$$

Par conséquent,

$$\begin{aligned}
\frac{1}{n}\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}) &= \frac{N_{UV}}{n}\, C + 2\sum_{d=1}^{h-3} \delta(\mathbb{W}; d)\, \frac{1}{n}\mathbb{E}(N_{\mathbb{W}(d)\mathbb{W}} \mid \mathbb{S}) - \frac{N_{UV}}{n}\, C^2(2d_{UV} + 1) \\
&+ \frac{N_{UV}^2}{n^2}\, C^2 \left(\sum_A \frac{n}{N_{A+}} d_{A+}^2 - \sum_{AB\ comp} \frac{n}{N_{AB}} d_{AB}^2\right) + \mathrm{O}(\frac{1}{n})\,.
\end{aligned}$$

On utilise maintenant le fait que $C \xrightarrow{ps} \dfrac{\mu(\mathbb{W})}{\mu(UV)}$, $\dfrac{1}{n}\mathbb{E}(N_{\mathbb{W}^{(d)}\mathbb{W}} \mid \mathbb{S}) \xrightarrow{ps} \mu(\mathbb{W}^{(d)}\mathbb{W})$, et que si AB est compatible, $\dfrac{N_{AB}}{n} \xrightarrow{ps} \mu(AB)$; on obtient

$$\begin{aligned}
\lim_{n \longrightarrow \infty} \frac{1}{n}\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}) &= \mu(\mathbb{W}) + 2\sum_{d=1}^{h-3} \delta(\mathbb{W}; d)\, \mu(\mathbb{W}^{(d)}\mathbb{W}) - \frac{\mu(\mathbb{W})^2}{\mu(UV)}(2d_{UV}+1) \\
&+ \mu(\mathbb{W})^2 \left(\sum_A \frac{d_{A+}^2}{\mu(A)} - \sum_{AB\ comp} \frac{d_{AB}^2}{\mu(AB)} \right) \quad p.s. \qquad (2.5)
\end{aligned}$$

- Si UV n'est pas compatible,

$$\begin{aligned}
&\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}) - 2\sum_{d=1}^{h-3} \delta(\mathbb{W}; d)\, \mathbb{E}(N_{\mathbb{W}^{(d)}\mathbb{W}} \mid \mathbb{S}) \\
&\quad = N_{U+}\, C + \frac{N_{U+}^2}{n}\, C^2(\Delta_2 - 2\Delta_1) + \mathrm{O}(1) \\
&\quad = N_{U+}\, C + \frac{N_{U+}^2}{n}\, C^2 \left(\sum_A \frac{n}{N_{A+}} d_{A+}^2 - \sum_{AB\ comp} \frac{n}{N_{AB}} d_{AB}^2 \right) + \mathrm{O}(1)\,.
\end{aligned}$$

Par conséquent,

$$\begin{aligned}
\frac{1}{n}\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}) &= \frac{N_{U+}}{n}\, C + 2\sum_{d=1}^{h-3} \delta(\mathbb{W}; d)\, \frac{1}{n}\mathbb{E}(N_{\mathbb{W}^{(d)}\mathbb{W}} \mid \mathbb{S}) \\
&+ \frac{N_{U+}^2}{n^2}\, C^2 \left(\sum_A \frac{n}{N_{A+}} d_{A+}^2 - \sum_{AB\ comp} \frac{n}{N_{AB}} d_{AB}^2 \right) + \mathrm{O}(\frac{1}{n})\,.
\end{aligned}$$

Or $\dfrac{N_{U+}}{n} \xrightarrow{ps} \mu(U)$ et $C \xrightarrow{ps} \dfrac{\mu(\mathbb{W})}{\mu(U)}$, ce qui donne

$$\begin{aligned}
\lim_{n \to \infty} \frac{1}{n}\mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}) &= \mu(\mathbb{W}) + 2\sum_{d=1}^{h-3} \delta(\mathbb{W}; d)\mu(\mathbb{W}^{(d)}\mathbb{W}) \\
&+ \mu(\mathbb{W})^2 \left(\sum_A \frac{d_{A+}^2}{\mu(A)} - \sum_{AB\ comp} \frac{d_{AB}^2}{\mu(AB)} \right) \quad p.s. \qquad (2.6)
\end{aligned}$$

En utilisant les définitions $d_{AB} = n_{AB}$ si $AB \neq UV$ et $d_{UV} = n_{UV} - 1$, on s'aperçoit que les deux formules (2.5) et (2.6) sont identiques. En effet,

- si UV est compatible alors

$$2d_{UV} + 1 = 2n_{UV} - 1\,,$$

$$\begin{aligned}\sum_{AB\ comp}\frac{d_{AB}^2}{\mu(AB)} &= \sum_{AB\neq UV,\,AB\ comp}\frac{n_{AB}^2}{\mu(AB)}+\frac{(n_{UV}-1)^2}{\mu(UV)}\\ &= \sum_{AB\ comp}\frac{n_{AB}^2}{\mu(AB)}+\frac{(1-2n_{UV})}{\mu(UV)}\ ;\end{aligned}$$

- si UV n'est pas compatible alors

$$\sum_{AB\ comp}\frac{d_{AB}^2}{\mu(AB)}=\sum_{AB\ comp}\frac{n_{AB}^2}{\mu(AB)}.$$

Dans les deux cas, on a :

$$\begin{aligned}\sum_{A}\frac{d_{A+}^2}{\mu(A)} &= \sum_{A}\frac{(\sum_B d_{AB})^2}{\mu(A)}\\ &= \sum_{A\neq U}\frac{(\sum_B d_{AB})^2}{\mu(A)}+\frac{(\sum_B d_{UB})^2}{\mu(U)}\\ &= \sum_{A\neq U}\frac{n_{A+}^2}{\mu(A)}+\frac{(\sum_B n_{UB}-1)^2}{\mu(U)}\\ &= \sum_{A}\frac{n_{A+}^2}{\mu(A)}+\frac{1-2n_{U+}}{\mu(U)}.\end{aligned}$$

Par conséquent, on a presque sûrement

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathrm{Var}(N_{\mathbb{W}}\mid\mathbb{S}) &= \mu(\mathbb{W})+2\sum_{d=1}^{h-3}\delta(\mathbb{W};d)\mu(\mathbb{W}^{(d)}\mathbb{W})\\ &+ \mu(\mathbb{W})^2\left(\sum_{A}\frac{n_{A+}^2}{\mu(A)}+\frac{1-2n_{U+}}{\mu(U)}-\sum_{AB\ comp}\frac{n_{AB}^2}{\mu(AB)}\right).\end{aligned}\qquad(2.7)$$

On établit alors le lemme suivant :

Lemme 3 : Variance conditionnelle asymptotique dans *M2*

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}Var(N_W\mid\mathcal{S}_2) &= \mu(W)+2\sum_{d=1}^{h-3}\delta(W;d)\,\mu(W^{(d)}W)\\ &+ \mu(W)^2\left(\sum_{a,b}\frac{n_{ab+}^2}{\mu(ab)}-\sum_{a,b,c}\frac{n_{abc}^2}{\mu(abc)}+\frac{1-2n_{w_1w_2+}}{\mu(w_1w_2)}\right)\quad p.s.\end{aligned}$$

où n_{abc} est le nombre d'occurrences de abc dans le mot W et $n_{ab+}=\sum_c n_{abc}$.

Preuve : il suffit de reprendre la formule (2.7) et d'écrire $\mathbb{W}$, A, U, AB (compatible) dans l'alphabet $\mathcal{A}$. □

En Annexe A.2, on vérifie que pour les mots de longueur 4, $w_1w_2w_3w_4$, l'expression asymptotique donnée par le Lemme 3, est bien la même que celle de la variance asymptotique de $n^{-1/2}\left(N(w_1w_2w_3w_4) - \dfrac{N(w_1w_2w_3)N(w_2w_3w_4)}{N(w_2w_3)}\right)$ obtenue par une approche basée sur des propriétés de martingales (voir Chapitre 3, formule (3.7)), et définie par

$$\sigma^2(w_1w_2w_3w_4) = \frac{\mu(w_1w_2w_3w_4)}{\mu(w_2w_3)^2}\left(\mu(w_2w_3) - \mu(w_2w_3w_4)\right)\left(\mu(w_2w_3) - \mu(w_1w_2w_3)\right). \qquad (2.8)$$

En effet, ces deux approches, sont équivalentes dans le modèle *Mm* d'ordre m pour les mots de longueur $h = m + 2$.
Cette écriture assure que la variance asymptotique est, d'une part positive, et d'autre part strictement positive dès lors que toutes les transitions du mot $W = w_1w_2w_3w_4$ sont non nulles et que la dernière de ces transitions, $\pi(w_2w_3, w_4)$, est différente de 1.

2.4 Généralisation au modèle Mm

La méthode que l'on vient de détailler dans le modèle *M2* se généralise facilement pour les chaînes de Markov d'ordre m (*Mm*). En effet, il suffit de relier l'ensemble $\mathcal{S}_m$ des séquences sur l'alphabet $\mathcal{A}$ ayant leurs m premières lettres et leurs comptages $N(a_1 \cdots a_{m+1})$ de $(m+1)$-mots fixés, avec l'ensemble $\mathbb{S}_m$ des séquences sur l'alphabet $\mathcal{A}^m$ ayant leur première lettre fixée, ainsi que les comptages N_{AB} de leurs 2-mots AB définis comme

$$N_{AB} = \begin{cases} N(a_1a_2\cdots a_m a_{m+1}) & \text{si } A = a_1\cdots a_m \text{ et } B = a_2\cdots a_{m+1} (AB \text{ compatible}), \\ 0 & \text{sinon } (AB \text{ non compatible}). \end{cases}$$

De plus, un mot $W = w_1w_2\cdots w_h$ de longueur h dans l'alphabet $\mathcal{A}$ correspond au mot $\mathbb{W} = W_1W_2\cdots W_{h-m+1}$ de longueur $h - m + 1$ dans l'alphabet $\mathcal{A}^m$ tel que $W_i = w_iw_{i+1}\cdots w_{i+m-1}$.
Les résultats suivants restent vrais

$$\begin{aligned} \mathbb{E}(N_W \mid \mathcal{S}_m) &= \mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}_m) = n\widehat{\mu}_m(W) + \mathrm{O}(1) \quad p.s. \\ \mathrm{Var}(N_W \mid \mathcal{S}_m) &= \mathrm{Var}(N_{\mathbb{W}} \mid \mathbb{S}_m), \end{aligned}$$

où $n\widehat{\mu}_m(W)$ est l'estimateur naturel de $n\mu_m(W)$:

$$n\widehat{\mu}_m(W) = \frac{\prod_{j=1}^{h-m} N(w_jw_{j+1}\cdots w_{j+m})}{\prod_{j=2}^{h-m} N(w_jw_{j+1}\cdots w_{j+m-1})}.$$

Le théorème suivant s'obtient ainsi comme le Théorème 1.

Théorème 3
$$U_m(W) = \frac{N_W - \mathbb{E}(N_W \mid \mathcal{S}_m)}{Var(N_W \mid \mathcal{S}_m)^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$
La statistique $U_m(W)$ est bien définie pour n assez grand.

La formule asymptotique de $\frac{1}{n}\mathrm{Var}(N_W \mid \mathcal{S}_m)$ est donnée par le lemme suivant, et se démontre comme au Lemme 3.

Lemme 4 : Variance conditionnelle asymptotique dans *Mm*

$$\lim_{n\to\infty} \frac{1}{n} Var(N_W \mid \mathcal{S}_m) = \mu(W) \ + 2 \sum_{d=1}^{h-m-1} \delta(W;d)\, \mu(W^{(d)}W)$$
$$+ \mu(W)^2 \left(\sum_{a_1,\dots,a_m} \frac{n^2_{a_1\cdots a_m +}}{\mu(a_1\cdots a_m)} - \sum_{a_1,\dots,a_{m+1}} \frac{n^2_{a_1\cdots a_m a_{m+1}}}{\mu(a_1\cdots a_m a_{m+1})} + \frac{1-2n_{w_1\cdots w_m +}}{\mu(w_1\cdots w_m)} \right) \quad p.s.$$

Remarques

- Si m vaut 0 dans le membre de droite de l'équation du Lemme 4, et en prenant soin d'écrire que $N_{a_1\cdots a_m+} = \sum_{a_{m+1}} N_{a_1\cdots a_{m+1}}$ et $\mu(a_1\cdots a_m) = \sum_{a_{m+1}} \mu(a_1\cdots a_{m+1})$, on obtient exactement l'expression asymptotique de $\frac{1}{n}\mathrm{Var}(N_W \mid \mathcal{S}_0)$ (donnée par la formule (A.8) en Annexe), expression qui correspond au modèle *M0* dans lequel les lettres X_i sont supposées indépendantes.
- Il est intéressant de remarquer que l'on retrouve les résultats de Prum *et al.* (1995) en faisant $m = 1$ dans les formules de l'espérance et la variance conditionnelles asymptotiques ci-dessus.
- Lorsque $m = h-2$, l'expression asymptotique donnée par le Lemme 4 est identique à celle de la variance asymptotique de $n^{-1/2}(N_W - n\widehat{\mu}_m(W))$ obtenue par l'approche martingale du Chapitre 3. Pour un mot $W = w_1\cdots w_h$, la formule (3.7) s'écrit

$$\sigma^2_W = \frac{\mu(W)}{\mu(w_2\cdots w_{h-1})^2} \left(\mu(w_2\cdots w_{h-1}) - \mu(w_2\cdots w_h)\right)\left(\mu(w_2\cdots w_{h-1}) - \mu(w_1\cdots w_{h-1})\right).$$

Chapitre 3

Modèle markovien d'ordre m et de période 3

3.1 Introduction

Lorsqu'une séquence d'ADN est codante, elle est constituée d'une succession de trinucléotides, appelés codons, dont la propriété est de coder pour des acides aminés. Les acides aminés sont au nombre de 20 et sont codés par 61 codons. Ainsi, plusieurs codons codent pour le même acide aminé ; les deux premières bases du codon déterminent souvent l'acide aminé. Les trois codons restant, appelés "codons stop", signalent la fin du message génétique (gène) à traduire en protéine (succession d'acides aminés). Le *cadre de lecture* détermine le "découpage" de la séquence en codons. Un oligonucléotide (mot) peut donc apparaître sur trois positions différentes par rapport au cadre de lecture, et coder ainsi plusieurs suites d'acides aminés suivant cette position. Nous appellerons *phase* cette position par rapport au cadre de lecture. Il paraît donc naturel de différencier les occurrences d'un mot en phase 1, 2 ou 3. Pour des commodités d'écriture, nous définissons les conventions suivantes : la longueur d'une séquence est multiple de 3 et sa première lettre est le début d'un codon ; on dira que le mot $W = w_1 w_2 \cdots w_h$ apparaît en position i dans la séquence $S = X_1 X_2 \cdots X_n$ si $(X_{i-h+1}, X_{i-h+2}, \ldots, X_{i-1}, X_i) = (w_1, w_2, \ldots, w_{h-1}, w_h)$, et qu'il est en phase k si de plus, $i = k$[modulo 3] (on écrira par la suite $i\%3 = k$).

L'objectif de ce chapitre est d'une part, de modéliser une séquence d'ADN codante, et d'autre part d'étudier dans ce modèle le nombre d'occurrences d'un mot W en phase k (k =1, 2 ou 3), ou toutes phases confondues. Soient $W = w_1 w_2 \cdots w_h$ un mot de longueur h et $k \in \{1, 2, 3\}$; on définit le nombre $N_{W,k}$ d'occurrences de W en phase k dans la séquence S par

$$N_{W,k} = \sum_{i=h}^{n} \mathbb{I}\{X_{i-h+1} X_{i-h+2} \cdots X_i = w_1 w_2 \cdots w_h \text{ et } i\%3 = k\}$$

$$N_{W,k} = \sum_{\substack{j \\ h \le 3j+k \le n}} \mathbb{1}\{\text{le mot } W \text{ termine en position } 3j+k \text{ dans } S\}\,.$$

Comme au Chapitre 2, on cherchera à définir à partir de $N_{W,k}$, des statistiques asymptotiquement gaussiennes standard.

Modélisation

• Un premier modèle consiste à choisir indépendamment les X_i parmi $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ avec la loi de probabilité $\mu_k = (\mu(\mathtt{A}, k), \mu(\mathtt{C}, k), \mu(\mathtt{G}, k), \mu(\mathtt{T}, k))$ si $i\%3 = k$; pour les séquences observées, dès que la séquence dépasse le millier de paires de bases, un tel modèle s'y ajuste très mal. Ainsi, on est amené à considérer des modèles de chaînes de Markov qui définissent des classes de modèles dépendant de plus de paramètres.

• Considérons, dans un premier temps, le modèle dans lequel la séquence S est générée par une chaîne de Markov d'ordre 1 périodique de période 3 (modèle que l'on notera *M1_3*), c'est à dire dont les transitions se définissent ainsi :

$$\forall a, b \in \mathcal{A},\ \forall k \in \{1,2,3\},\ \mathrm{P}\{X_{3j+k} = b \mid X_{3j+k-1} = a\} = \pi_k(a,b)\,,$$

où $\Pi_k = (\pi_k(a,b))_{a,b\in\mathcal{A}}$, $k \in \{1,2,3\}$, sont trois matrices de transition telles que les éléments $\pi_k(a,b) > 0\,, \forall a, b \in \mathcal{A},\ \forall k \in \{1,2,3\}$.
Une telle chaîne peut se paramétrer de deux façons pour obtenir une chaîne de Markov homogène d'ordre 1.

- La première consiste à écrire la séquence S dans l'alphabet à 12 lettres $\overline{\mathcal{A}} = \mathcal{A} \times \{1,2,3\}$. On obtient ainsi la séquence $\overline{S} = \overline{X}_1\overline{X}_2\cdots\overline{X}_n$ où $\overline{X}_i = (X_i, i\%3) \in \overline{\mathcal{A}}$. $\overline{S}$ est une chaîne de Markov dont la matrice de transition $\overline{\Pi} = (\overline{\pi}(\overline{a},\overline{b}))_{\overline{a},\overline{b}\in\overline{\mathcal{A}}}$ est définie par

  $$\overline{\pi}(\overline{a},\overline{b}) = \begin{cases} \pi_k(a,b) & \text{si } \overline{a} = (a, k-1) \text{ et } \overline{b} = (b,k), \\ 0 & \text{sinon .} \end{cases}$$

  Cette formalisation est plaisante pour les calculs mais $\overline{\Pi}$ a des termes nuls et la chaîne $\overline{S}$ est périodique de période 3. L'utilisation des théorèmes ergodiques, du type loi des grands nombres ou théorème de la limite centrale, n'est donc pas immédiat. Cependant, on va voir ci-dessous qu'ils sont valides en paramétrisant S autrement.

- La deuxième paramétrisation consiste à grouper les X_i trois par trois, c'est à dire considérer une chaîne de Markov d'ordre 1 stationnaire particulière dans l'alphabet $\mathcal{A}^3$: on pose $X'_j = X_{3j-2}X_{3j-1}X_{3j}$ et on écrit la séquence S comme $S' = X'_1X'_2\cdots X'_{n'}$ avec $n = 3n'$. La matrice de transition Π' correspondante est telle que

  $$\pi'(abc, def) = \pi_1(c,d)\,\pi_2(d,e)\,\pi_3(e,f) > 0,$$

  où les transitions π_k ont été définies auparavant. La chaîne de Markov S' est donc ergodique et on note μ' l'unique mesure invariante sur $\mathcal{A}^3$.

Un mot $W = w_1 w_2 \cdots w_h$ en phase k dans la séquence S correspond à un mot phasé, noté (W, k), dans $\overline{S}$ qui s'écrit dans l'alphabet $\overline{\mathcal{A}}$ de la façon suivante :

$$(W, k) = (w_1, (k - h + 1)\%3) \cdots (w_{h-1}, (k - 1)\%3)(w_h, k).$$

Le nombre d'occurrences de (W, k) dans $\overline{S}$ peut s'exprimer comme une somme de nombre d'occurrences dans S' de mots sur l'alphabet $\mathcal{A}^3$. Par exemple,

$$N_{abcd,1} = \sum_{e,f \in \mathcal{A}} N_{abcdef,3} = \sum_{B'=def,\, e,f \in \mathcal{A}} N_{A'B'}(S') \quad \text{avec } A' = abc.$$

Par conséquent, en utilisant l'ergodicité de la chaîne S', on obtient des théorèmes limites pour la chaîne $\overline{S}$. En effet, plaçons-nous dans un cas simple à décrire, celui pour lequel $h = 3h'$ et $k = 3$. Ainsi, en définissant $W'_i = w_i w_{i+1} w_{i+2}$ et $W' = W'_1 W'_2 \cdots W'_{h'}$, on a $N_{W,3} = N_{W'}(S')$. Par la loi des grands nombres

$$\frac{1}{n'} N_{W'}(S') \xrightarrow{ps} \mu'(W'),$$

d'où $\frac{1}{n} N_{W,3} \xrightarrow{ps} \frac{1}{3}\mu'(W')$. Or,

$$\begin{aligned}
\mathbb{E}(N_{W,3}) &= \sum_{j=h'}^{n'} \mathrm{P}\{(W,3) \text{ termine en } 3j \text{ dans } \overline{S}\} \\
&= \sum_{j=h'}^{n'} \mathrm{P}\{W' \text{ termine en } j \text{ dans } S'\} \\
&= (n' - h' + 1)\mu'(W') = \frac{n - h + 3}{3}\mu'(W')\ ;
\end{aligned}$$

par conséquent, en posant

$$\mu(W, 3) = \frac{1}{3}\mu'(W') = \frac{1}{3}\ \mathrm{P}\{(W,3) \text{ termine en } 3j \text{ dans } \overline{S}\},$$

on a $\mathbb{E}(N_{W,3}) = (n - h + 3)\mu(W, 3)$ et $\frac{1}{n} N_{W,3} \xrightarrow{ps} \mu(W, 3)$.

Dans le cas général, en définissant

$$\mu(W, k) = \frac{1}{3}\ \mathrm{P}\{(W,k) \text{ termine en } 3j + k \text{ dans } \overline{S}\},$$

on a

$$\frac{1}{n} N_{W,k} \xrightarrow{ps} \mu(W, k).$$

Remarque : le facteur $\frac{1}{3}$ introduit dans la définition de $\mu(W, k)$ permet de garder la formule habituelle $\mathbb{E}(N_{W,k}) \simeq n\mu(W, k)$ quand on utilise l'alphabet $\overline{\mathcal{A}}$ et de considérer toujours $N_{W,k}$ comme somme de n indicatrices, même si deux sur trois sont automatiquement nulles.

• On suppose maintenant que S est une chaîne de Markov d'ordre m dont les transitions dépendent de la phase (modèle *Mm_3*) : $\forall a_1, a_2, \dots a_{m+1} \in \mathcal{A}$, $\forall k \in \{1,2,3\}$,

$$\mathrm{P}\{X_{3j+k} = a_{m+1} \mid X_{3j+k-1} = a_m, \dots, X_{3j+k-m} = a_1\} = \pi_k^{(m)}(a_1 \cdots a_m, a_{m+1}) > 0.$$

La chaîne $S' = X_1' X_2' \cdots X_{n'}'$ définie un peu plus haut est une chaîne de Markov sur $\mathcal{A}^3$ ergodique d'ordre m', où m' est la partie entière supérieure de $m/3$. Il suffit ensuite de construire la chaîne de Markov d'ordre 1 $S'' = X_1'' X_2'' \cdots X_{n'-m'+1}''$ où $X_\ell'' = X_\ell' X_{\ell+1}' \cdots X_{\ell+m'-1}' \in \mathcal{A}^{3m'}$, comme au Chapitre 2. Cette chaîne S'' est alors ergodique et permet de faire hériter la chaîne $\overline{S}$ sur $\overline{\mathcal{A}}$ de cette propriété, et les théorèmes limites (LGN, TLC) s'appliquent.

Pour deux types d'estimateurs consistants de $n\mu(W,k)$, nous proposons une normalisation de la statistique d'écart

$$T(W,k) = \frac{1}{\sqrt{n}} \left(N_{W,k} - n\widehat{\mu}(W,k) \right)$$

telle que cette statistique normalisée soit asymptotiquement distribuée suivant une loi Normale centrée réduite. Dans un premier temps, nous estimons $n\mu(W,k)$ par le compensateur de $N_{W,k}$, ce qui nous permet d'utiliser un théorème de la limite centrale pour les martingales (Section 3.2). Ensuite, nous estimons $n\mu(W,k)$ par l'espérance de $N_{W,k}$ conditionnellement à la statistique exhaustive du modèle et nous utilisons les techniques développées au Chapitre 2 (Section 3.3). Ces deux approches ont été introduite par Prum *et al.* (1995) dans le cadre d'une chaîne de Markov homogène d'ordre 1 (modèle *M1*).

3.2 Approche martingale dans le modèle Mm_3

On se place dans le modèle *Mm_3* , où $1 \leq m \leq h-2$ et $h \in \mathbb{N}^*$ est fixé, c'est à dire que l'on suppose que la séquence $S = X_1 X_2 \cdots X_n$ est générée par une chaîne de Markov d'ordre m, dont les transitions sont définies par

$$\mathrm{P}\{X_{3j+k} = a_{m+1} \mid X_{3j+k-1} = a_m, \dots, X_{3j+k-m} = a_1\} = \pi_k^{(m)}(a_1 \cdots a_m, a_{m+1}) > 0\,,$$

où les $a_i \in \mathcal{A}$ et $k \in \{1,2,3\}$. Pour alléger les notations, nous écrirons π_k au lieu de $\pi_k^{(m)}$.

On considère la filtration naturelle associée à $(X_i)_{i=1\cdots n}$, c'est à dire $\mathcal{F} = (\mathcal{F}_i)_{i=1\cdots n}$ où $\mathcal{F}_i$ est la σ-algèbre engendrée par $X_1, \dots, X_i$.

3.2.1 Étude du comptage $N_{W,k}$ d'un mot W en phase k

Soient $W = w_1 \cdots w_h$ un mot de longueur h et $k \in \{1,2,3\}$; $\mu(W,k)$ définie en Section 3.1 peut s'écrire

$$\mu(W,k) = \mu(W^-, k-1)\pi_k(\widetilde{W}^-, w_h)\,,$$

avec les notations suivantes :

$$\begin{aligned}
W^- &= w_1 \cdots w_{h-1} \text{ (le mot } W \text{ privé de sa dernière lettre)},\\
\widetilde{W} &= w_{m-h} \cdots w_h \text{ (le mot composé des } m+1 \text{ dernières lettres de } W),\\
\widetilde{W}^- &= w_{m-h} \cdots w_{h-1} \text{ (le mot } \widetilde{W} \text{ privé de sa dernière lettre)} .
\end{aligned}$$

En définissant les variables

$$\begin{aligned}
Z_i &= 1\!\!\mathrm{I}\{\text{un mot } W \text{ se termine en position } i \text{ dans } S\},\\
\alpha_i^{(k)} &= 1\!\!\mathrm{I}\{i\%3 = k\},
\end{aligned}$$

le nombre $N_{W,k}^{(n)}$ d'occurrences du mot W en phase k dans S s'écrit

$$N_{W,k}^{(n)} = \sum_{i=h}^{n} \alpha_i^{(k)} Z_i \; .$$

Le compensateur de la suite $(N_{W,k}^{(n)})_{n\in\mathbb{N}}$ est la suite $\left(\sum_{i=h}^{n} \mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1})\right)_{n\in\mathbb{N}}$. Par conséquent $(M_n)_{n\in\mathbb{N}} = \left(N_{W,k}^{(n)} - \sum_{i=h}^{n} \mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1})\right)_{n\in\mathbb{N}}$ est une martingale adaptée à $\mathcal{F}$.
De façon plus explicite, on a

$$\begin{aligned}
\mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1}) &= \alpha_i^{(k)} \mathbb{E}(Z_i \mid \mathcal{F}_{i-1})\\
&= \alpha_i^{(k)} \mathrm{P}\{W \text{ termine en } i \mid X_1, \ldots, X_{i-1}\}\\
&= \alpha_i^{(k)} 1\!\!\mathrm{I}\{W^- \text{ termine en } i-1\} \Pi_{i\%3}(\widetilde{W}^-, w_h). \qquad (3.1)
\end{aligned}$$

Par conséquent, M_n s'écrit

$$M_n = N_{W,k} - N_{W^-,k-1} \pi_k(\widetilde{W}^-, w_h) \; .$$

Soit $< M >_n$ le crochet de la martingale M_n

$$< M >_n = \sum_{i=h}^{n} \mathbb{E}([\alpha_i^{(k)} Z_i - \mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1})]^2 \mid \mathcal{F}_{i-1}) \; ;$$

un théorème de convergence pour martingales vectorielles (Duflo, 1990, théorème 3.II.11), nous garantit la convergence en loi de $\frac{1}{\sqrt{n}} M_n$ vers une gaussienne centrée sous les hypothèses suivantes :

Théorème 4 *Soit M_n une martingale vectorielle de carré intégrable adaptée à une filtration $(\mathcal{F}_n)$; on note $< M >_n$ son crochet.*
On suppose que les deux hypothèses suivantes sont réalisées :
(i) $\frac{1}{n} < M >_n \xrightarrow{P} \Gamma$;
(ii) la condition de Lindeberg est satisfaite :

$$\forall \varepsilon > 0, \; \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(||M_i - M_{i-1}||^2 1\!\!\mathrm{I}\{||M_i - M_{i-1}|| \geq \varepsilon\sqrt{n}\} \mid \mathcal{F}_{i-1}) \xrightarrow{P} 0 \; .$$

Alors,

$$\frac{1}{\sqrt{n}} M_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma).$$

Le calcul de $< M >_n$ nous donne

$$\begin{aligned} < M >_n &= \sum_{i=h}^{n} \mathbb{E}\left([\alpha_i^{(k)} Z_i - \mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1})]^2 \mid \mathcal{F}_{i-1} \right) \\ &= \sum_{i=h}^{n} [\mathbb{E}\left([\alpha_i^{(k)} Z_i]^2 \mid \mathcal{F}_{i-1} \right) - \mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1})^2] \\ &= \sum_{i=h}^{n} [\mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1}) - \mathbb{E}(\alpha_i^{(k)} Z_i \mid \mathcal{F}_{i-1})^2] \\ &= N_{W^-,k-1} \pi_k(\widetilde{W}^-, w_h)[1 - \pi_k(\widetilde{W}^-, w_h)] \text{, en utilisant (3.1).} \end{aligned}$$

Par conséquent, la loi forte des grands nombres pour $N_{W^-,k-1}$ donne

$$\lim_{n \to +\infty} \frac{1}{n} < M >_n = \mu(W^-, k-1) \pi_k(\widetilde{W}^-, w_h)[1 - \pi_k(\widetilde{W}^-, w_h)] \, .$$

La condition de Lindeberg étant vérifiée, on obtient

$$\frac{1}{\sqrt{n}} \left(N_{W,k} - N_{W^-,k-1} \pi_k(\widetilde{W}^-, w_h) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

avec

$$\sigma^2 = \mu(W^-, k-1) \pi_k(\widetilde{W}^-, w_h)[1 - \pi_k(\widetilde{W}^-, w_h)] \, . \tag{3.2}$$

On considère alors la statistique

$$T(W,k) = \frac{1}{\sqrt{n}} \left(N_{W,k} - N_{W^-,k-1} \widehat{\pi}_k(\widetilde{W}^-, w_h) \right)$$

où $\widehat{\pi}_k$ est l'estimateur empirique de π_k :

$$\widehat{\pi}_k(\widetilde{W}^-, w_h) = \frac{N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}} \, ,$$

soit

$$T(W,k) = \frac{1}{\sqrt{n}} \left(N_{W,k} - \frac{N_{W^-,k-1} N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}} \right) \, .$$

$\frac{N_{W^-,k-1} N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}}$ est un estimateur consistant de $n\mu(W,k) = n\mu(W^-, k-1)\pi_k(\widetilde{W}^-, w_h)$ car on a la convergence presque sûre de $\frac{N_{W^-,k-1}}{n}$ vers $\mu(W^-, k-1)$.

On va voir que $T(W,k)$ converge vers une gaussienne centrée dont on précisera la variance. Pour cela, on écrit $T(W,k)$ de la manière suivante :

$$T(W,k) = \frac{1}{\sqrt{n}} \left(N_{W,k} - N_{W^-,k-1} \pi_k(\widetilde{W}^-, w_h) \right) - \frac{1}{\sqrt{n}} \frac{N_{W^-,k-1}}{N_{\widetilde{W}^-,k-1}} \left(N_{\widetilde{W},k} - N_{\widetilde{W}^-,k-1} \pi_k(\widetilde{W}^-, w_h) \right)$$

afin d'appliquer le Théorème 4 à la martingale vectorielle $(A_{W,k}, B_{W,k})_{n\in\mathbb{N}}$ où

$$\begin{aligned} A_{W,k} &= N_{W,k} - N_{W^-,k-1}\pi_k(\widetilde{W}^-, w_h), &(3.3)\\ B_{W,k} &= N_{\widetilde{W},k} - N_{\widetilde{W}^-,k-1}\pi_k(\widetilde{W}^-, w_h)\,. &(3.4)\end{aligned}$$

On obtient ainsi

$$\left(\frac{1}{\sqrt{n}}A_{W,k}, \frac{1}{\sqrt{n}}B_{W,k}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{(k)})$$

avec

$$\Sigma^{(k)} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix} \tag{3.5}$$

où, d'après (3.2)

$$\begin{aligned} \sigma_{11}^2 &= \mu(W^-, k-1)\pi_k(\widetilde{W}^-, w_h)[1-\pi_k(\widetilde{W}^-, w_h)]\,,\\ \sigma_{22}^2 &= \mu(\widetilde{W}^-, k-1)\pi_k(\widetilde{W}^-, w_h)[1-\pi_k(\widetilde{W}^-, w_h)]\,,\end{aligned}$$

et

$$\begin{aligned} \sigma_{12} &= \lim_{n\to+\infty}\frac{1}{n}\sum_{i=h}^{n}\mathbb{E}\left(\left[\alpha_i^{(k)}Z_i - \mathbb{E}(\alpha_i^{(k)}Z_i \mid \mathcal{F}_{i-1})\right]\left[\alpha_i^{(k)}\widetilde{Z}_i - \mathbb{E}(\alpha_i^{(k)}\widetilde{Z}_i|\mathcal{F}_{i-1})\right] \mid \mathcal{F}_{i-1}\right)\\ &\quad \text{avec}\quad \widetilde{Z}_i = \mathbb{1}\{\widetilde{W} \text{ termine en } i \text{ dans } S\},\\ &= \lim_{n\to+\infty}\frac{1}{n}\sum_{i=h}^{n}\mathbb{E}\left([\alpha_i^{(k)}Z_i - \mathbb{E}(\alpha_i^{(k)}Z_i \mid \mathcal{F}_{i-1})]^2 \mid \mathcal{F}_{i-1}\right),\\ &= \sigma_{11}^2\,.\end{aligned}$$

Par conséquent, on a

$$T(W,k) = \frac{1}{\sqrt{n}}\left(N_{W,k} - \frac{N_{W^-,k-1}N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{W,k}^2) \tag{3.6}$$

avec

$$\begin{aligned} \sigma_{W,k}^2 &= \sigma_{11}^2\left(1 - 2\frac{\mu(W^-,k-1)}{\mu(\widetilde{W}^-,k-1)}\right) + \left(\frac{\mu(W^-,k-1)}{\mu(\widetilde{W}^-,k-1)}\right)^2\sigma_{22}^2\\ &= \pi_k(\widetilde{W}^-, w_h)[1-\pi_k(\widetilde{W}^-, w_h)]\left(\mu(W^-,k-1) - \frac{\mu(W^-,k-1)^2}{\mu(\widetilde{W}^-,k-1)}\right)\\ &= \pi_k(\widetilde{W}^-, w_h)[1-\pi_k(\widetilde{W}^-, w_h)]\mu(W^-,k-1)\left(1 - \frac{\mu(W^-,k-1)}{\mu(\widetilde{W}^-,k-1)}\right)\\ &= \frac{\mu(W,k)}{\mu(\widetilde{W}^-,k-1)}\Big(\mu(W,k) - \mu(W^-,k-1) - \mu(\widetilde{W},k) + \mu(\widetilde{W}^-,k-1)\Big)\\ &= \frac{\mu(W,k)}{\mu(\widetilde{W}^-,k-1)^2}\Big(\mu(\widetilde{W}^-,k-1) - \mu(\widetilde{W},k)\Big)\Big(\mu(\widetilde{W}^-,k-1) - \mu(W^-,k-1)\Big)\,.\end{aligned} \tag{3.7}$$

Afin de normaliser la statistique $T(W,k)$, on choisit un estimateur consistant de $\sigma^2_{W,k}$, par exemple

$$\widehat{\sigma}^2_{W,k} = \frac{N_{W^-,k-1} N_{\widetilde{W},k}}{n\, N^3_{\widetilde{W}^-,k-1}} \left(N_{\widetilde{W}^-,k-1} - N_{W^-,k-1}\right)\left(N_{\widetilde{W}^-,k-1} - N_{\widetilde{W},k}\right) , \tag{3.8}$$

et on obtient le théorème suivant :

Théorème 5 *La statistique*

$$U(W,k) = \frac{T(W,k)}{\widehat{\sigma}_{W,k}} = \frac{n^{-1/2}}{\widehat{\sigma}_{W,k}} \left(N_{W,k} - \frac{N_{W^-,k-1} N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}}\right)$$

converge en loi vers la loi Normale centrée réduite.
$\widehat{\sigma}^2_{W,k}$ est ici un estimateur consistant de $\sigma^2_{W,k}$, par exemple celui donné par la formule (3.8).

Cette statistique, correspondant à un mot fixé (longueur et phase fixées), convient pour tester si $\mu(W,k) = \mu(W^-, k-1)\pi_k(\widetilde{W}^-, w_h)$, c'est à dire si le fait d'ajouter la lettre w_h après un mot W^- en phase $k-1$ est compatible avec le modèle *Mm_3* . En effet, $U(W,k)$ est la statistique normalisée calculée à partir de l'écart entre les rapports $\frac{N_{W,k}}{N_{W^-,k-1}}$ et $\frac{N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}}$. Ainsi, à l'aide de la statistique $U(W,k)$, on peut déterminer si le mot phasé (W,k) a une fréquence "exceptionnelle" ou non dans la séquence. Le caractère exceptionnel proviendra de la succession du mot W^- en phase $k-1$ et de la lettre w_h en phase k.

3.2.2 Étude du comptage N_W d'un mot W

En utilisant le fait que le nombre N_W d'occurrences d'un mot W dans la séquence S s'écrit

$$N_W = N_{W,1} + N_{W,2} + N_{W,3} ,$$

on peut obtenir une nouvelle statistique $T(W)$ construite à partir de N_W, asymptotiquement gaussienne centrée (dont on précisera la variance) et en déduire une statistique $U(W)$ asymptotiquement gaussienne standard.

Pour cela, on applique le Théorème 4 à la martingale vectorielle $(A_{W,1}, B_{W,1}, A_{W,2}, B_{W,2}, A_{W,3}, B_{W,3})$, dont les expressions de chaque composante sont données par (3.3) et (3.4), et on obtient

$$\frac{1}{\sqrt{n}}(A_{W,1}, B_{W,1}, A_{W,2}, B_{W,2}, A_{W,3}, B_{W,3}) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\Sigma)$$

avec

$$\Sigma = \begin{pmatrix} \Sigma^{(1)} & 0 & 0 \\ 0 & \Sigma^{(2)} & 0 \\ 0 & 0 & \Sigma^{(3)} \end{pmatrix}$$

où $\Sigma^{(1)}$, $\Sigma^{(2)}$ et $\Sigma^{(3)}$ sont définies par (3.5). Σ est diagonale par bloc car les termes de covariance correspondant aux couples $(A_{W,\ell}, A_{W,\ell'})$, $(A_{W,\ell}, B_{W,\ell'})$ et $(B_{W,\ell}, B_{W,\ell'})$ sont nuls dès que $\ell \neq \ell'$;

en effet, l'élément de Σ sur la ligne 3 et la colonne 1 (correspondant au couple $(A_{W,1}, A_{W,2})$), par exemple, est

$$\Sigma_{13} = \lim_{n\to+\infty} \frac{1}{n} \sum_{i=h}^{n} \alpha_i^1 \alpha_i^2 \mathbb{E}\left([Z_i - \mathbb{E}(Z_i \mid \mathcal{F}_{i-1})]^2 \mid \mathcal{F}_{i-1}\right) = 0$$

car $\alpha_i^1 \alpha_i^2 = 0$.
On obtient ainsi la convergence suivante

$$T(W) = \frac{1}{\sqrt{n}} \left(N_W - \sum_{k=1}^{3} \frac{N_{W^-,k-1} N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_W^2)$$

où

$$\sigma_W^2 = \sigma_{W,1}^2 + \sigma_{W,2}^2 + \sigma_{W,3}^2 \ ,$$

et $\sigma_{W,1}^2$, $\sigma_{W,2}^2$ et $\sigma_{W,3}^2$ sont données par la formule (3.7).

Pour normaliser $T(W)$, il suffit de prendre un estimateur $\widehat{\sigma}_W^2$ consistant de σ_W^2, par exemple

$$\widehat{\sigma}_W^2 = \widehat{\sigma}_{W,1}^2 + \widehat{\sigma}_{W,2}^2 + \widehat{\sigma}_{W,3}^2 \tag{3.9}$$

où $\widehat{\sigma}_{W,1}^2$, $\widehat{\sigma}_{W,2}^2$ et $\widehat{\sigma}_{W,3}^2$ sont donnés par la formule (3.8). On a alors le théorème suivant :

Théorème 6 *La statistique*

$$U(W) = \frac{T(W)}{\widehat{\sigma}_W} = \frac{n^{-1/2}}{\widehat{\sigma}_W} \left(N_W - \sum_{k=1}^{3} \frac{N_{W^-,k-1} N_{\widetilde{W},k}}{N_{\widetilde{W}^-,k-1}} \right)$$

converge en loi vers la loi Normale centrée réduite. $\widehat{\sigma}_W^2$ est ici un estimateur consistant de σ_W^2, par exemple celui donné par la formule (3.9).

Cette statistique globale permet cette fois de comparer le nombre d'occurrences d'un mot, toutes phases confondues, avec celui attendu dans le modèle *Mm_3* et calculé en regard du mot privé de sa dernière lettre.

L'utilisation des quatre statistiques $U(W,1)$, $U(W,2)$, $U(W,3)$ et $U(W)$ pour savoir si, par exemple, un mot a une fréquence exceptionnelle dans S, doit être faite en parallèle car chacune apporte une information différente et complémentaire.

3.3 Approche conditionnelle dans le modèle Mm_3

Nous allons étudier en détail le modèle *M1_3* , puis nous généraliserons aisément les résultats au modèle général *Mm_3* , en se basant sur le travail effectué au Chapitre 2 qui avait comme objectif d'adapter, au modèle *Mm* , la méthode conditionnelle développée par Prum *et al.* (1995) dans *M1*.

Soit $S = X_1 X_2 \cdots X_n$ une séquence générée par une chaîne de Markov d'ordre 1 sur l'alphabet $\mathcal{A}$, dont les transitions dépendent de la phase :

$$\forall a, b \in \mathcal{A},\ \forall k \in \{1, 2, 3\},\ \mathrm{P}\{X_{3j+k} = b \mid X_{3j+k-1} = a\} = \pi_k(a, b) > 0\,.$$

On écrit cette séquence dans l'alphabet $\overline{\mathcal{A}} = \mathcal{A} \times \{1, 2, 3\}$ comme cela est décrit en Section 3.1. La séquence $\overline{S} = \overline{X}_1 \overline{X}_2 \cdots \overline{X}_n$, où $\overline{X}_i = (X_i, i\%3)$, ainsi formée est une chaîne de Markov d'ordre 1 ergodique de matrice de transition $\overline{\Pi} = (\overline{\pi}(\overline{a}, \overline{b}))_{\overline{a}, \overline{b} \in \overline{\mathcal{A}}}$ telle que

$$\overline{\pi}((a, \ell), (b, k)) = \pi_k(a, b) \mathbb{1}\{k = (\ell + 1)\%3\}.$$

Si on représente par une matrice $\overline{N}$, l'ensemble des comptages de 2-mots $\overline{a}\overline{b}$ dans la séquence $\overline{S}$, on a :

$$\overline{N} = \begin{pmatrix} 0 & N_2 & 0 \\ 0 & 0 & N_3 \\ N_1 & 0 & 0 \end{pmatrix},$$

où N_ℓ, $\ell \in \{1, 2, 3\}$, est la matrice de dimension 4 des comptages $N_{ab,\ell}$ des 2-mots ab en phase ℓ dans la séquence S :

$$N_\ell = \begin{pmatrix} N_{\mathtt{AA},\ell} & N_{\mathtt{AC},\ell} & N_{\mathtt{AG},\ell} & N_{\mathtt{AT},\ell} \\ N_{\mathtt{CA},\ell} & N_{\mathtt{CC},\ell} & N_{\mathtt{CG},\ell} & N_{\mathtt{CT},\ell} \\ N_{\mathtt{GA},\ell} & N_{\mathtt{GC},\ell} & N_{\mathtt{GG},\ell} & N_{\mathtt{GT},\ell} \\ N_{\mathtt{TA},\ell} & N_{\mathtt{TC},\ell} & N_{\mathtt{TG},\ell} & N_{\mathtt{TT},\ell} \end{pmatrix}.$$

Nous utiliserons les notations suivantes :

$$\begin{aligned} \overline{N}_{\overline{a}+} &= \sum_{\overline{b}} \overline{N}_{\overline{a}\overline{b}} \\ N_{a+,\ell} &= \sum_{b} N_{ab,\ell}\ ; \end{aligned}$$

on remarquera que si $\overline{a} = (a, \ell - 1)$, alors $\overline{N}_{\overline{a}+} = N_{a+,\ell}$, et si $\overline{b} = (b, \ell)$ alors $\overline{a}\overline{b} = \overline{ab} = (ab, \ell)$.

Soit $\mathbb{S}$ l'ensemble de toutes les séquences de longueur n sur l'alphabet $\overline{\mathcal{A}}$, commençant par $\overline{X}_1$ et ayant $\overline{N}(\overline{S})$ comme matrice de comptages des 2-mots ; par abus de notation, on définit la matrice $\overline{N}(\mathbb{S}) = \overline{N}(\overline{S}) = \overline{N}$.

Notre but est d'établir la normalité asymptotique standard de la statistique conditionnelle

$$U_{1\_3}(W, k) = \frac{N_{W,k} - \mathbb{E}(N_{W,k} \mid \mathbb{S})}{\sqrt{\mathrm{Var}(N_{W,k} \mid \mathbb{S})}}$$

pour un mot W de longueur h et une phase k fixés. On notera souvent $\overline{W}$ pour le couple (W,k) afin de rappeler que l'on travaille avec l'alphabet $\overline{\mathcal{A}}$, et, par soucis de clarté, $N_{\overline{W}}$ désignera le nombre d'occurrences de $\overline{W}$ dans $\overline{S}$ (on évite ainsi les notations $\overline{N}_{\overline{W}}(\overline{S})$ et $\overline{N}_{\overline{W}}$).

Dans un premier temps, nous calculons l'espérance de $N_{\overline{W}}$ conditionnellement à $\mathbb{S}$ en utilisant les techniques du Chapitre 2. Ce calcul est direct lorsque la longueur h du mot W est égale à deux modulo trois car la délétion de l'intérieur du mot $\overline{W}$ ne crée alors pas de rupture de phase dans la séquence $\overline{S}\setminus\overline{W}$. Il en est autrement si h n'est pas de la forme $3p+2$. On devra ainsi distinguer trois cas. Le calcul de la variance conditionnelle s'obtiendra également en considérant les trois cas successifs, et fera apparaître des termes de covariances conditionnelles entre deux mots différents $\overline{W}$ et $\overline{W'}$. Les covariances seront aussi utilisées pour calculer la variance conditionnelle de $N_W = N_{W,1} + N_{W,2} + N_{W,3}$. Enfin, nous énoncerons les résultats asymptotiques obtenus.

3.3.1 Espérance conditionnelle de $N_{W,k}$ dans M1_3

Soient $W = w_1 w_2 \cdots w_h$ et $k \in \{1,2,3\}$; pour simplifier les écritures, on pose $u = w_1$ et $v = w_h$.
On considère successivement les trois cas : $h = 3p+2$, $h = 3p+1$ et $h = 3p$.

Mots de longueur $3p+2$

À partir d'une séquence $\mathbf{S} \in \mathbb{S}$ dans laquelle un mot $(W,k) = \overline{W}$ apparaît (termine) en i,

$$\begin{array}{rcccccccccccccc} \mathbf{S} & = & x_1 & x_2 & x_3 & x_4 & \cdots & u & w_2 & \cdots & w_{h-1} & v & x_{i+1} & \cdots & x_n \\ & & \uparrow & \uparrow & \uparrow & \uparrow & & \uparrow & \uparrow & & \uparrow & \uparrow & \uparrow & & \uparrow \\ & & 1 & 2 & 3 & 1 & \cdots & (k-1) & k & \cdots & k-1 & k & (k+1) & \cdots & n\%3 \end{array}$$

on construit la séquence $\mathbf{S}\setminus\overline{W}$ dans laquelle le mot $\overline{W}$ a été substitué par le mot (uv,k) :

$$\begin{array}{rcccccccccccc} \mathbf{S}\setminus\overline{W} & = & x_1 & x_2 & x_3 & x_4 & \cdots & u & v & x_{i+1} & \cdots & x_n\ ; \\ & & \uparrow & \uparrow & \uparrow & \uparrow & & \uparrow & \uparrow & \uparrow & & \uparrow \\ & & 1 & 2 & 3 & 1 & \cdots & (k-1) & k & (k+1) & \cdots & n\%3 \end{array}$$

le fait d'enlever "l'intérieur" du mot $\overline{W}$ ne perturbe pas les phases suivantes.
Ainsi, la matrice $\overline{N}(\mathbf{S}\setminus\overline{W})$ des comptages de 2-mots associée est de même structure que $\overline{N}(\mathbb{S})$, c'est à dire avec trois blocs non nuls $N_\ell(\mathbf{S}\setminus\overline{W})$ tels que

$$N_{ab,\ell}(\mathbf{S}\setminus\overline{W}) = N_{ab,\ell}(\mathbb{S}) - N_{ab,\ell}(\overline{W}) + \mathbb{1}\{(a,b,\ell) = (u,v,k)\}\,,$$

où $N_{ab,\ell}(\overline{W})$ est le nombre d'occurrences de (ab,ℓ) dans (W,k) ; $\overline{N}(\overline{W})$ désignera la matrice de dimension 12 des comptages de 2-mots dans $\overline{W}$.

Soit $\mathbb{S}\setminus\overline{W}$ l'ensemble de toutes les séquences associées à la matrice $\overline{N}(\mathbf{S}\setminus\overline{W})$ et commençant par $\overline{X}_1$. On établit ainsi une bijection entre les séquences de $\mathbb{S}$ ayant un mot $\overline{W}$ en position i, et les séquences de $\mathbb{S} \setminus \overline{W}$ ayant un mot $\overline{uv} = (uv, k)$ en position $i - h + 2$; ces deux ensembles ont donc des cardinaux égaux, notés $\beta_i(\overline{W}, \mathbb{S})$ et $\beta_{i-h+2}(\overline{uv}, \mathbb{S} \setminus \overline{W})$.

Les séquences de $\mathbb{S}$ étant équiprobables de probabilité $(\#\mathbb{S})^{-1}$, on écrit donc les relations suivantes :

$$\begin{aligned}
\mathbb{E}(N_{\overline{W}} \mid \mathbb{S}) &= \sum_{i=h}^{n} \mathrm{P}\{\overline{W} \text{ termine en position } i \mid \mathbb{S}\} \\
&= \frac{1}{\#\mathbb{S}} \sum_{i=h}^{n} \beta_i(\overline{W}, \mathbb{S}) \\
&= \frac{1}{\#\mathbb{S}} \sum_{i=h}^{n} \beta_{i-h+2}(\overline{uv}, \mathbb{S} \setminus \overline{W}).
\end{aligned}$$

De façon analogue, on a l'équation suivante

$$\mathbb{E}(N_{\overline{uv}} \mid \mathbb{S} \setminus \overline{W}) = \frac{1}{\#\mathbb{S} \setminus \overline{W}} \sum_{i=2}^{n-h+2} \beta_i(\overline{uv}, \mathbb{S} \setminus \overline{W}) \ ;$$

or, $\mathbb{E}(N_{\overline{uv}} \mid \mathbb{S} \setminus \overline{W})$ est déterministe et vaut $N_{\overline{uv}}(\mathbb{S} \setminus \overline{W})$. Par conséquent, on obtient

$$\mathbb{E}(N_{\overline{W}} \mid \mathbb{S}) = N_{\overline{uv}}(\mathbb{S} \setminus \overline{W}) \ \frac{\#\mathbb{S} \setminus \overline{W}}{\#\mathbb{S}}. \tag{3.10}$$

Chaque cardinal est calculé avec la formule de Whittle (1955) ci-dessous

$$\#\mathbb{S} = K(\mathbb{S}) H(\mathbb{S}) \tag{3.11}$$

où

$$K(\mathbb{S}) = \prod_{\overline{a}} \frac{\overline{N}_{\overline{a}+}!}{\prod_{\overline{b}} \overline{N}_{\overline{a}\overline{b}}!} = \prod_{a,\ell} \frac{N_{a+,\ell}!}{\prod_b N_{ab,\ell}!}$$

et $H(\mathbb{S})$ est le cofacteur dans la matrice $I - \widehat{\overline{\Pi}}$ de n'importe quel élément de la ligne correspondant à la lettre $\overline{X}_n$; $\widehat{\overline{\Pi}}$ est la matrice de transition empirique associée à $\mathbb{S}$, c'est à dire $\widehat{\overline{\pi}}(\overline{a}, \overline{b}) = \overline{N}_{\overline{a}\overline{b}} / \overline{N}_{\overline{a}+}$, $\overline{a}, \overline{b} \in \overline{\mathcal{A}}$.
Autrement dit, $\widehat{\overline{\Pi}}$ est la matrice de dimension 12 de même structure que $\overline{N}$, dont les trois blocs non nuls sont du type $\widehat{\Pi}_\ell = (\widehat{\pi}_\ell(a,b))_{a,b}$ avec $\widehat{\pi}_\ell(a,b) = N_{ab,\ell}/N_{a+,\ell}$, $a, b \in \mathcal{A}$, $\ell \in \{1, 2, 3\}$.

Mots de longueur $3p + 1$

On ne peut pas appliquer directement la méthode précédente car une séquence comme $\mathbf{S} \setminus \overline{W}$ est déphasée à partir de la lettre v. Aussi, pour y remédier, on écrit

$$N_{\overline{W}} = N_{\mathtt{A}W,k} + N_{\mathtt{C}W,k} + N_{\mathtt{G}W,k} + N_{\mathtt{T}W,k} + \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\},$$

de sorte que les mots $\mathtt{A}W$, $\mathtt{C}W$, $\mathtt{G}W$ et $\mathtt{T}W$ sont à leur tour des mots de longueur $3p+2$; leur espérance conditionnelle se calcule avec la formule (3.10). On a toutefois besoin de calculer $\mathrm{P}\{\overline{S}$ commence par $\overline{W} \mid \mathbb{S}\}$. Il est clair que cette probabilité est nulle si $u \neq X_1$ ou si $k \neq 1$.

- Si $u = X_1$ et $k = h\%3 = 1$, alors à toute séquence $\mathbf{S} \in \mathbb{S}$ qui commence par $\overline{W}$

$$\begin{array}{ccccccccc} \mathbf{S} & = & u & w_2 & \cdots & v & x_{h+1} & x_{h+2} & \cdots & x_n \\ & & \uparrow & \uparrow & & \uparrow & \uparrow & \uparrow & & \uparrow \\ & & 1 & 2 & \cdots & 1 & 2 & 3 & \cdots & n\%3 \end{array}$$

  on associe la séquence $\widetilde{\mathbf{S}}^{\overline{W}}$ suivante qui commence par v :

$$\begin{array}{cccccccc} \widetilde{\mathbf{S}}^{\overline{W}} & = & v & x_{h+1} & x_{h+2} & \cdots & x_n \\ & & \uparrow & \uparrow & \uparrow & & \uparrow \\ & & 1 & 2 & 3 & \cdots & n\%3 \end{array}$$

  et dont la matrice des comptages de 2-mots est $\overline{N}(\widetilde{\mathbf{S}}^{\overline{W}}) = \overline{N}(\mathbb{S}) - \overline{N}(\overline{W})$. On note $\widetilde{\mathbb{S}}^{\overline{W}}$ l'ensemble des séquences associées à la matrice $\overline{N}(\widetilde{\mathbf{S}}^{\overline{W}})$ et commençant par v.

- Sinon, on pose $\widetilde{\mathbb{S}}^{\overline{W}} = \emptyset$.

Par conséquent, on a

$$\mathrm{P}\{\overline{S} \text{ commence par } \overline{W} \mid \mathbb{S}\} = \frac{\#\widetilde{\mathbb{S}}^{\overline{W}}}{\#\mathbb{S}} \tag{3.12}$$

où $\#\widetilde{\mathbb{S}}^{\overline{W}}$ se calcule par la formule de Whittle.

Ainsi,

$$\mathbb{E}(N_{\overline{W}} \mid \mathbb{S}) = \sum_a \mathbb{E}(N_{aW,k} \mid \mathbb{S}) + \frac{\#\widetilde{\mathbb{S}}^{\overline{W}}}{\#\mathbb{S}} . \tag{3.13}$$

où les formules d'espérances conditionnelles sont données par (3.10).

Mots de longueur $3p$

Ici, on écrit

$$N_{\overline{W}} = \sum_{a,b} N_{abW,k} \; + \; \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \; + \; \sum_a \mathbb{1}\{\overline{S} \text{ commence par } \overline{aW}\}$$

où $\overline{aW} = (aW, k)$. On en déduit donc

$$\mathbb{E}(N_{\overline{W}} \mid \mathbb{S}) = \sum_{a,b} \mathbb{E}(N_{abW,k} \mid \mathbb{S}) + \mathrm{P}\{\overline{S} \text{ commence par } \overline{W} \mid \mathbb{S}\} + \sum_a \mathrm{P}\{\overline{S} \text{ commence par } \overline{aW} \mid \mathbb{S}\} .$$

L'expression des espérances conditionnelles est donnée par (3.10) car abW est un mot de longueur $3p+2$; le dernier terme se calcule à partir de la formule (3.12) ainsi que le second, en utilisant la décomposition :

$$\mathrm{P}\{\overline{S} \text{ commence par } \overline{W} \mid \mathbb{S}\} = \sum_a \mathrm{P}\{\overline{S} \text{ commence par } \overline{Wa} \mid \mathbb{S}\}\,, \tag{3.14}$$

avec $\overline{Wa} = (Wa, k+1)$. Ainsi,

$$\mathbb{E}(N_{\overline{W}} \mid \mathbb{S}) = \sum_{a,b} \mathbb{E}(N_{abW,k} \mid \mathbb{S}) + \sum_a \frac{\#\widetilde{\mathbb{S}}^{\overline{Wa}}}{\#\mathbb{S}} + \sum_a \frac{\#\widetilde{\mathbb{S}}^{\overline{aW}}}{\#\mathbb{S}}\,. \tag{3.15}$$

3.3.2 Variance et covariance conditionnelles dans M1_3

On adapte la méthode décrite au paragraphe précédent pour calculer $\mathrm{Var}(N_{\overline{W}} \mid \mathbb{S})$. Nous calculerons aussi les covariances conditionnelles du type $\mathrm{Cov}(N_{W,k}, N_{W',k'} \mid \mathbb{S})$ pour deux mots W et W' de longueur $h = 3p+2$.

Mots de longueur $3p+2$: calcul de la variance conditionnelle de $N_{\overline{W}}$

On veut calculer

$$\mathbb{E}(N^2_{\overline{W}} \mid \mathbb{S}) = \sum_{i,i'} \mathrm{P}\{\text{deux mots } \overline{W} \text{ terminent en } i \text{ et en } i' \mid \mathbb{S}\}$$

On développe cette somme en trois termes :

- un terme T_1 correspondant à $i = i'$.
- un terme T_2 correspondant à $|i - i'| \geq h - 1$. Celui-ci correspond au cas où deux mots $\overline{W}$ apparaissent de façon non chevauchante dans la séquence (on autorise toutefois le chevauchement d'une seule lettre, qui n'est possible que si $u = v$).
- un terme T_3 correspondant à $|i - i'| < h - 1$. Celui-ci correspond au cas où deux mots $\overline{W}$ peuvent se chevaucher sur plus d'une lettre.

– T_1, obtenu en faisant $i = i'$, vaut : $T_1 = \mathbb{E}(N_{\overline{W}} \mid \mathbb{S})$.
– Le second terme est

$$T_2 = 2 \sum_{\substack{i < i' \\ |i-i'| \geq h-1}} \mathrm{P}\{\text{deux mots } \overline{W} \text{ terminent en } i \text{ et en } i' \mid \mathbb{S}\}\,;$$

comme pour le calcul de l'espérance, on écrit

$$T_2 = \frac{2}{\#\mathbb{S}} \sum_{\substack{i < i' \\ |i-i'| \geq h-1}} \beta_{i,i'}(\overline{W}, \mathbb{S})$$

où $\beta_{i,i'}(\overline{W},\mathbb{S})$ est le nombre de séquences de $\mathbb{S}$ dans lesquelles le mot $\overline{W}$ apparaît à la fois aux positions i et i'. On utilise alors la bijection entre ces séquences et celles obtenues en enlevant l'intérieur des deux mots $\overline{W}$ en i et i'. Ces dernières appartiennent à l'ensemble $\mathbb{S}\setminus\overline{W}\setminus\overline{W}$ des séquences dont les comptages de (ab,ℓ) sont les $N_{ab,\ell}$ corrigés par $2(N_{ab,\ell}(\overline{W}) - \mathbb{1}\{(ab,\ell)=(uv,k)\})$, et possèdent les mots $\overline{uv}$ aux positions $i-h+2$ et $i'-2h+4$. En notant $\beta_{i-h+2,i'-2h+4}(\overline{uv},\mathbb{S}\setminus\overline{W}\setminus\overline{W})$ le nombre de telles séquences, on a par bijection

$$\beta_{i,i'}(\overline{W},\mathbb{S}) = \beta_{i-h+2,i'-2h+4}(\overline{uv},\mathbb{S}\setminus\overline{W}\setminus\overline{W})\ .$$

En utilisant le fait que le nombre de paires de $\overline{uv}$ dans $\mathbb{S}\setminus\overline{W}\setminus\overline{W}$ est déterministe et vaut

$$\frac{N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}\setminus\overline{W})(N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}\setminus\overline{W})-1)}{2}$$

on obtient

$$T_2 = N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}\setminus\overline{W})(N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}\setminus\overline{W})-1)\ \frac{\#\mathbb{S}\setminus\overline{W}\setminus\overline{W}}{\#\mathbb{S}}\ .$$

– Le dernier terme

$$T_3 = 2\sum_{d=1}^{h-2}\ \sum_{\substack{i<i'\\ i'-i=d}} \mathrm{P}\{\text{deux mots } \overline{W} \text{ terminent en } i \text{ et en } i' \mid \mathbb{S}\}\ ;$$

correspond au cas où pour un d tel que $1\le d\le h-2$, les $h-d$ premières lettres de $\overline{W}$ sont identiques aux $h-d$ dernières, d'où la possibilité d'avoir le mot $\overline{W}$ en position i et $i+d$. On introduit les notations suivantes : si le chevauchement est possible, ce qui implique que $d\%3=0$, on pose $\delta(\overline{W};d)=1$ et on note $\overline{W}^{(d)}\overline{W}=(w_1\cdots w_d w_1\cdots w_h,k)$ le mot obtenu en concaténant les d premières lettres de $\overline{W}$ avec $\overline{W}$; sinon $\delta(\overline{W};d)=0$.
Ainsi, T_3 s'écrit

$$\begin{aligned} T_3 &= 2\sum_{d=1}^{h-2}\delta(\overline{W};d)\mathbb{E}(N_{\overline{W}^{(d)}\overline{W}}\mid\mathbb{S})\\ &= 2\sum_{d=1}^{h-2}\delta(\overline{W};d)N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}^{(d)}\overline{W})\ \frac{\#\mathbb{S}\setminus\overline{W}^{(d)}\overline{W}}{\#\mathbb{S}}\ . \end{aligned}$$

On a donc le résultat suivant :

$$\begin{aligned} \mathbb{E}(N_{\overline{W}}^2\mid\mathbb{S}) &= N_{\overline{uv}}(\mathbb{S}\setminus\overline{W})\ \frac{\#\mathbb{S}\setminus\overline{W}}{\#\mathbb{S}}\\ &+ N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}\setminus\overline{W})(N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}\setminus\overline{W})-1)\ \frac{\#\mathbb{S}\setminus\overline{W}\setminus\overline{W}}{\#\mathbb{S}}\\ &+ 2\sum_{d=1}^{h-2}\delta(\overline{W};d)N_{\overline{uv}}(\mathbb{S}\setminus\overline{W}^{(d)}\overline{W})\ \frac{\#\mathbb{S}\setminus\overline{W}^{(d)}\overline{W}}{\#\mathbb{S}} \end{aligned} \tag{3.16}$$

où chaque cardinal se calcule par la formule de Whittle.

On en déduit l'expression de la variance conditionnelle :

$$\mathrm{Var}(N_{\overline{W}} \mid \mathbb{S}) = \mathbb{E}(N_{\overline{W}}^2 \mid \mathbb{S}) - \mathbb{E}(N_{\overline{W}} \mid \mathbb{S})^2 \ .$$

Mots de longueur $3p+2$: calcul des covariances conditionnelles

Pour calculer la variance conditionnelle de $N_{\overline{W}}$ quand $h \neq 3p+2$, nous avons besoin de l'expression de $\mathrm{Cov}(N_{W,k}, N_{W',k} \mid \mathbb{S})$ où W et W' sont deux mots de longueur $3p+2$; par ailleurs, pour calculer la variance conditionnelle de $N_W = N_{W,1} + N_{W,2} + N_{W,3}$, on utilisera des termes du type $\mathrm{Cov}(N_{W,k}, N_{W,k'} \mid \mathbb{S})$ où W est un mot de longueur $3p+2$. C'est pourquoi, nous calculons de façon générale $\mathrm{Cov}(N_{W,k}, N_{W',k'} \mid \mathbb{S})$ avec $W = u \cdots v$ et $W' = u' \cdots v'$ deux mots de longueur $h = 3p+2$, k et k' sont deux phases appartenant à $\{1,2,3\}$. Nous noterons $\overline{W} = (W,k)$ et $\overline{W'} = (W', k')$.

Le calcul de $\mathbb{E}(N_{\overline{W}} N_{\overline{W'}} \mid \mathbb{S})$ se fait de façon analogue à celui de $\mathbb{E}(N_{\overline{W}}^2 \mid \mathbb{S})$. On a l'expression suivante :

$$\mathbb{E}(N_{\overline{W}} N_{\overline{W'}} \mid \mathbb{S}) = \sum_{i,i'} \mathrm{P}\{\overline{W} \text{ termine en } i \text{ et } \overline{W'} \text{ termine en } i' \mid \mathbb{S}\} .$$

On développe cette somme en trois termes :

- un terme T'_1 correspondant à $i = i'$. Celui-ci sera nul dès que $\overline{W} \neq \overline{W'}$.
- un terme T'_2 correspondant à $|i - i'| \geq h - 1$. Celui-ci correspond au cas où les mots $\overline{W}$ et $\overline{W'}$ apparaissent de façon non chevauchante dans la séquence (on autorise toutefois le chevauchement d'une seule lettre, qui n'est possible que si $(u, (k-h+1)\%3) = (v', k')$ ou $(u', (k'-h+1)\%3) = (v, k)$).
- un terme T'_3 correspondant à $|i - i'| < h - 1$. Celui-ci correspond au cas où les deux mots $\overline{W}$ et $\overline{W'}$ se chevauchent sur plus d'une lettre.

Si le mot $\overline{W}$ en position i est compatible avec le mot $\overline{W'}$ en position $i + d$, on note $\overline{W}^{(d)}\overline{W'}$ ce mot obtenu par chevauchement et $\delta(\overline{W}; \overline{W'}; d) = 1$. Sinon $\delta(\overline{W}; \overline{W'}; d) = 0$.

– T'_1, obtenu en faisant $i = i'$, vaut $T'_1 = \mathbb{1}\{\overline{W} = \overline{W'}\}\mathbb{E}(N_{\overline{W}} \mid \mathbb{S})$.
– Le second terme vaut

$$T'_2 = N_{\overline{uv}}(\mathbb{S} \setminus \overline{W} \setminus \overline{W'}) \left(N_{\overline{u'v'}}(\mathbb{S} \setminus \overline{W} \setminus \overline{W'}) - \delta(\overline{uv}; \overline{u'v'}; 0)\right) \frac{\#\mathbb{S} \setminus \overline{W} \setminus \overline{W'}}{\#\mathbb{S}} ,$$

où $\overline{uv} = (uv, k)$ et $\overline{u'v'} = (u'v', k')$.

– Le dernier terme, s'écrit

$$T'_3 = \sum_{d \in \{2-h, \dots, h-2\} \setminus \{0\}} \delta(\overline{W}; \overline{W}'; d) \mathbb{E}(N_{\overline{W}^{(d)}\overline{W'}}) \mid \mathbb{S}) .$$

En remarquant que T'_1 peut s'écrire

$$T'_1 = \delta(\overline{W}; \overline{W'}; 0) \mathbb{E}(N_{\overline{W}^{(0)}\overline{W'}} \mid \mathbb{S}) ,$$

on obtient la formule suivante :

$$\begin{aligned} \mathbb{E}(N_{\overline{W}} N_{\overline{W'}} \mid \mathbb{S}) = & \sum_{d=2-h}^{h-2} \delta(\overline{W}; \overline{W}'; d) \mathbb{E}(N_{\overline{W}^{(d)}\overline{W'}} \mid \mathbb{S}) \\ + \quad & N_{\overline{uv}}(\mathbb{S} \setminus \overline{W} \setminus \overline{W'}) (N_{\overline{u'v'}}(\mathbb{S} \setminus \overline{W} \setminus \overline{W'}) - \delta(\overline{uv}; \overline{u'v'}; 0)) \, \frac{\#\mathbb{S} \setminus \overline{W} \setminus \overline{W'}}{\#\mathbb{S}} . \end{aligned} \tag{3.17}$$

On en déduit l'expression de la covariance :

$$\mathrm{Cov}(N_{\overline{W}} N_{\overline{W'}} \mid \mathbb{S}) = \mathbb{E}(N_{\overline{W}} N_{\overline{W'}} \mid \mathbb{S}) - \mathbb{E}(N_{\overline{W}} \mid \mathbb{S}) \, \mathbb{E}(N_{\overline{W'}} \mid \mathbb{S}) .$$

Mots de longueur $3p+1$

À partir de l'égalité suivante,

$$N_{\overline{W}} = \sum_a N_{aW,k} + \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\}$$

on a

$$\begin{aligned} \mathbb{E}(N^2_{\overline{W}} \mid \mathbb{S}) \quad = \quad & \sum_{a,b} \mathbb{E}(N_{aW,k} N_{bW,k} \mid \mathbb{S}) \\ + \quad & \mathrm{P}\{\overline{S} \text{ commence par } \overline{W} \mid \mathbb{S}\} \\ + \quad & 2 \sum_a \mathbb{E}(N_{aW,k} \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S}) , \end{aligned}$$

où les deux premiers termes se calculent avec (3.17) et (3.12) ; l'expression de $\mathbb{E}(N_{aW,1} \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S})$ est donnée par (B.1) en Annexe.

On en déduit l'expression de la variance :

$$\mathrm{Var}(N_{\overline{W}} \mid \mathbb{S}) = \mathbb{E}(N^2_{\overline{W}} \mid \mathbb{S}) - \mathbb{E}(N_{\overline{W}} \mid \mathbb{S})^2 .$$

Mots de longueur $3p$

Pour se ramener à des mots de longueur $3p+2$, on écrit

$$N_{\overline{W}} = \sum_{a,b} N_{abW,k} + \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} + \sum_c \mathbb{1}\{\overline{S} \text{ commence par } \overline{cW}\} ,$$

où $\overline{cW} = (cW, k)$.
L'expression de $\mathbb{E}(N^2_{\overline{W}} \mid \mathbb{S})$ est

$$\begin{aligned}
\mathbb{E}(N^2_{\overline{W}} \mid \mathbb{S}) &= \sum_{a,b,a',b'} \mathbb{E}(N_{abW,k} N_{a'b'W,k} \mid \mathbb{S}) \\
&+ \mathrm{P}\{\overline{S} \text{ commence par } \overline{W} \mid \mathbb{S}\} \\
&+ 2\sum_{a,b} \mathbb{E}(N_{abW,k} \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S}) \\
&+ \sum_{c} \mathrm{P}\{\overline{S} \text{ commence par } \overline{cW} \mid \mathbb{S}\} \\
&+ 2\sum_{a,b,c} \mathbb{E}(N_{abW,k} \mathbb{1}\{\overline{S} \text{ commence par } \overline{cW}\} \mid \mathbb{S})\,.
\end{aligned}$$

Le premier terme se calcule à l'aide de (3.17), les formules (3.14) et (3.12) permettent de calculer le deuxième et le cinquième terme ; les expressions de $\mathbb{E}(N_{abW,3} \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S})$ et $\mathbb{E}(N_{abW,1} \mathbb{1}\{\overline{S} \text{ commence par } \overline{cW}\} \mid \mathbb{S})$ s'obtiennent à partir de (B.1), en remarquant que

$$\mathbb{E}(N_{abW,3} \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S}) = \sum_{c} \mathbb{E}(N_{abW,3} \mathbb{1}\{\overline{S} \text{ commence par } \overline{Wc}\} \mid \mathbb{S})\,,$$

où $\overline{Wc} = (Wc, k+1)$.

On en déduit l'expression de la variance :

$$\mathrm{Var}(N_{\overline{W}} \mid \mathbb{S}) = \mathbb{E}(N^2_{\overline{W}} \mid \mathbb{S}) - \mathbb{E}(N_{\overline{W}} \mid \mathbb{S})^2\,.$$

3.3.3 Espérance et variance conditionnelles de N_W dans M1_3

Comme dans l'approche martingale (Section **3.2**), il est aussi intéressant de calculer $\mathbb{E}(N_W \mid \mathbb{S})$ et $\mathrm{Var}(N_W \mid \mathbb{S})$ où N_W représente le nombre d'occurrences du mot W dans la séquence S.

On écrit

$$N_W = N_{W,1} + N_{W,2} + N_{W,3}\,, \tag{3.18}$$

d'où

$$\mathbb{E}(N_W \mid \mathbb{S}) = \sum_{k=1}^{3} \mathbb{E}(N_{W,k} \mid \mathbb{S})\,; \tag{3.19}$$

et on utilise l'une des formules (3.10), (3.13), (3.15) pour calculer les termes d'espérance.

Si $h = 3p+2$, on utilise la décomposition (3.18) :

$$\mathrm{Var}(N_W \mid \mathbb{S}) = \sum_{k=1}^{3} \mathrm{Var}(N_{\overline{W}} \mid \mathbb{S}) + 2\sum_{k<k'} \mathrm{Cov}(N_{W,k}, N_{W,k'} \mid \mathbb{S})$$

où les termes de variance se calculent à l'aide des formules (3.16) et (3.10), et les termes de covariance s'obtiennent en utilisant les formules (3.10) et (3.17) (en remplaçant W' par W).

Si $h=3p+1$ ou $h=3p$, on se ramène à des mots de longueur $h=3p+2$ en décomposant N_W de la façon suivante :

$$\begin{aligned} N_W &= \sum_a N_{aW} + \mathbb{1}\{\overline{S} \text{ commence par } (W,1)\}\,, \quad \text{si } h=3p+1 \\ N_W &= \sum_{a,b} N_{abW} + \mathbb{1}\{\overline{S} \text{ commence par } (W,3)\} + \sum_c \mathbb{1}\{\overline{S} \text{ commence par } (cW,1)\}\,, \\ & \qquad \text{si } h=3p\,. \end{aligned}$$

Ainsi, lorsque $h=3p+1$, le calcul de $\mathrm{Var}(N_W \,|\, \mathbb{S})$ utilise ceux de : $\mathrm{Var}(N_{aW} \,|\, \mathbb{S})$, $\mathrm{Cov}(N_{aW}, N_{bW} \,|\, \mathbb{S})$, $\mathrm{P}\{\overline{S} \text{ commence par } (W,1) \,|\, \mathbb{S}\}$ et $\mathrm{Cov}(N_{aW}, \mathbb{1}\{\overline{S} \text{ commence par } (W,1)\} \,|\, \mathbb{S})$;
de la même façon, lorsque $h=3p$, on a besoin de : $\mathrm{Var}(N_{abW} \,|\, \mathbb{S})$, $\mathrm{Cov}(N_{abW}, N_{a'b'W} \,|\, \mathbb{S})$, $\mathrm{P}\{\overline{S} \text{ commence par } (W,3) \,|\, \mathbb{S}\}$, $\mathrm{P}\{\overline{S} \text{ commence par } (cW,1) \,|\, \mathbb{S}\}$,
$\mathrm{Cov}(N_{abW}, \mathbb{1}\{\overline{S} \text{ commence par } (W,3)\} \,|\, \mathbb{S})$ et $\mathrm{Cov}(N_{abW}, \mathbb{1}\{\overline{S} \text{ commence par } (cW,1)\} \,|\, \mathbb{S})$.
Ces quantités sont calculées en Annexe **B**.

3.3.4 Résultats asymptotiques dans M1_3

Les statistiques T et U correspondant à cette approche sont

$$\begin{aligned} T_{1\_3}(W,k) &= \frac{1}{\sqrt{n}}(N_{W,k} - \mathbb{E}(N_{W,k} \,|\, \mathbb{S})) \\ U_{1\_3}(W,k) &= \frac{N_{W,k} - \mathbb{E}(N_{W,k} \,|\, \mathbb{S})}{\mathrm{Var}(N_{W,k} \,|\, \mathbb{S})^{1/2}} \end{aligned}$$

pour chaque phase, et

$$\begin{aligned} T_{1\_3}(W) &= \frac{1}{\sqrt{n}}(N_W - \mathbb{E}(N_W \,|\, \mathbb{S})) \\ U_{1\_3}(W) &= \frac{N_W - \mathbb{E}(N_W \,|\, \mathbb{S})}{\mathrm{Var}(N_W \,|\, \mathbb{S})^{1/2}} \end{aligned}$$

pour toutes phases confondues.

Le théorème suivant établit l'équivalence asymptotique entre $\mathbb{E}(N_{W,k} \,|\, \mathbb{S})$ et l'estimateur naturel de $n\mu(W,k)$, et le comportement asymptotique des statistiques $U_{1\_3}(W,k)$ centrées réduites.

Théorème 7 *Les statistiques $T_{1\_3}(W,k)$ sont telles que*

$$T_{1\_3}(W,k) - \frac{1}{\sqrt{n}}\left(N_{W,k} - n\widehat{\mu}_{1\_3}(W,k)\right) = O(n^{-1/2}) \quad p.s.$$

avec $n\widehat{\mu}_{1\_3}(W,k) = \dfrac{\prod_{j=1}^{h-1} N(w_j w_{j+1}, (k-h+j+1)\%3)}{\prod_{j=2}^{h-1} N(w_j, (k-h+j)\%3)}$ *, et les statistiques centrées réduites*

$$U_{1\_3}(W,k) = \frac{N_{W,k} - \mathbb{E}(N_{W,k} \,|\, \mathbb{S})}{Var(N_{W,k} \,|\, \mathbb{S})^{1/2}}$$

sont définies pour n assez grand et convergent en loi vers la loi Normale centrée réduite.

Preuve: Le schéma de la démonstration est rigoureusement identique à celui du Théorème 1 (Chapitre 2). L'espérance conditionnelle de $N_{W,k}$ est presque sûrement asymptotiquement équivalente à l'estimateur $n\widehat{\mu}_{1\_3}(W,k)$, et la normalité asymptotique découle d'un théorème de limite centrale pour $n^{-1/2}(N_{W,k} - n\widehat{\mu}_{1\_3}(W,k))$. Si la preuve du théorème peut se calquer parfaitement (aux phases près) sur celle du Théorème 1 lorsque le mot W est de longueur $3p+2$, l'existence d'une limite presque sûre de $\frac{1}{n}\mathrm{Var}(N_{W,k} \mid \mathbb{S})$ et la majoration du moment conditionnel d'ordre 4, dans le cas d'un mot de longueur $3p+1$ ou $3p$, nécessitent quelques développements techniques supplémentaires. Par exemple, si W est un mot de longueur $h = 3p+1$, on a

$$\begin{aligned}
\frac{1}{n}\mathrm{Var}(N_{W,k} \mid \mathbb{S}) &= \sum_{a,b} \frac{1}{n}\mathrm{Cov}(N_{aW,k}, N_{bW,k} \mid \mathbb{S}) \\
&+ \frac{1}{n}\mathrm{Var}(\mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S}) \\
&+ 2\sum_{a} \frac{1}{n}\mathrm{Cov}(N_{aW,k}, \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S}) .
\end{aligned}$$

Les deux derniers termes de ce développement tendent vers zéro quand n tend vers l'infini et, à titre d'exemple, nous calculons en Annexe **B** la limite de $\frac{1}{n}\mathrm{Cov}(N_{W,k}, N_{W',k'} \mid \mathbb{S})$ avec $|W| = |W'| = 3p+2$. □

Enfin, la convergence en loi de la statistique normalisée $U_{1\_3}(W)$ énoncée dans le théorème ci-dessous se démontre de la même façon en utilisant le fait que $N_W = N_{W,1} + N_{W,2} + N_{W,3}$.

Théorème 8 *La statistique centrée réduite*

$$U_{1\_3}(W) = \frac{N_W - \mathbb{E}(N_W \mid \mathbb{S})}{Var(N_W \mid \mathbb{S})^{1/2}}$$

est définie pour n assez grand et converge en loi vers la loi Normale centrée réduite.

Ces deux théorèmes nous permettrons dans la Partie III de détecter des mots ayant une fréquence exceptionnelle sur une phase donnée, ou bien toutes phases confondues, dans une séquence d'ADN. Ce caractère exceptionnel, relevant d'une fréquence significativement faible ou au contraire significativement élevée, sera quantifié par les statistiques U.

3.3.5 Généralisation dans Mm_3

Les résultats précédents se généralisent dans le modèle *Mm_3* de la même façon qu'au Chapitre 2, c'est à dire en écrivant la chaîne de Markov d'ordre m sur l'alphabet $\mathcal{A}$ comme une chaîne de Markov d'ordre 1 sur l'alphabet $\mathcal{A}^m$. Nous énonçons donc le théorème de convergence en loi des statistiques U_{m_3} associées.

Soit $\mathcal{S}_{m\_3}$ l'ensemble composé des m premières lettres de la séquence S, et des comptages $N(a_1a_2\cdots a_m,\ell)$ des m-mots $a_1a_2\cdots a_m$ en phase ℓ dans S.

Théorème 9 *Les statistiques centrées réduites*

$$U_{m\_3}(W,k)=\frac{N_{W,k}-\mathbb{E}(N_{W,k}\mid \mathcal{S}_{m\_3})}{Var(N_{W,k}\mid \mathcal{S}_{m\_3})^{1/2}}\quad et\quad U_{m\_3}(W)=\frac{N_W-\mathbb{E}(N_W\mid \mathcal{S}_{m\_3})}{Var(N_W\mid \mathcal{S}_{m\_3})^{1/2}}$$

sont définies pour n assez grand et convergent en loi vers la loi Normale centrée réduite.

L'espérance conditionnelle de $N_{W,k}$ est telle que

$$\mathbb{E}(N_{W,k}\mid \mathcal{S}_{m\_3})=n\widehat{\mu}_{m\_3}(W)\ +\ \mathrm{O}(1)\quad p.s.$$

avec

$$n\widehat{\mu}_{m\_3}(W)=\frac{\prod_{j=1}^{h-m}N(w_jw_{j+1}\cdots w_{j+m},(k-h+j+m)\%3)}{\prod_{j=2}^{h-m}N(w_jw_{j+1}\cdots w_{j+m-1},(k-h+j+m-1)\%3)}.$$

Si W est un mot de longueur $h=3p+m+1$, la variance asymptotique de $n^{-1/2}(N_{W,k}-\mathbb{E}(N_{W,k}\mid \mathcal{S}_{m\_3}))$ est donnée par

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathrm{Var}(N_{W,k}\mid \mathcal{S}_{m\_3})&=\mu(W,k)\ +2\sum_{d=1}^{h-m-1}\delta(\overline{W};d)\,\mu(W^{(d)}W,k)\\&+\mu(W,k)^2\left(\sum_{a_1,\dots,a_m,\ell}\frac{N^2_{(a_1\cdots a_m,\ell)+}(W,k)}{\mu(a_1\cdots a_m,\ell)}-\sum_{a_1,\dots,a_{m+1},\ell}\frac{N^2_{a_1\cdots a_ma_{m+1},\ell}(W,k)}{\mu(a_1\cdots a_ma_{m+1},\ell)}\right.\\&\left.\qquad+\frac{1-2N_{(w_1\cdots w_m,k-h+m)+}(W,k)}{\mu(w_1\cdots w_m,k-h+m)}\right)\quad p.s.\end{aligned}$$

Remarque : L'approche martingale et l'approche conditionnelle présentées dans ce chapitre sont équivalentes lorsque l'on se place dans le modèle d'ordre maximal, c'est à dire $m=h-2$ pour l'étude des mots de longueur h. En effet, si $m=h-2$

$$n\widehat{\mu}_{m\_3}(W)=\frac{N(w_1\cdots w_{h-1},k-1)N(w_2\cdots w_h,k)}{N(w_2\cdots w_{h-1},k)}$$

qui est précisément l'estimateur de $\mathbb{E}N_{W,k}$ du Théorème 5, et on peut vérifier que la limite presque sûre de $\frac{1}{n}\mathrm{Var}(N_{W,k}\mid \mathcal{S}_{m\_3})$ donnée ci-dessus est égale à la variance asymptotique donnée en (3.7).

Chapitre 4

Modèle markovien 3-périodique d'ordre 1 à codons fixés

4.1 Introduction

Le choix d'un modèle statistique pour une séquence d'ADN $X_1X_2\cdots X_n$ est délicat compte tenu des contraintes biologiques que l'on veut prendre en compte. En effet, les écarts au modèle choisi sont susceptibles de correspondre à des propriétés biologiques non encore identifiées.

Nous reprenons ici un modèle introduit par Altshul et Erickson (1985) qui s'appuie sur les deux constatations suivantes :

— il a été observé, et utilisé depuis quelques années, le fait que la fréquence des dinucléotides (2-mots) n'est pas compatible avec l'indépendance des bases X_i. Il est donc raisonnable de considérer des chaînes de Markov d'ordre faible comme modèles de base.

— La presque totalité du génome bactérien code pour des protéines qui sont des successions d'acides aminés. Ces derniers sont synthétisés directement à partir des codons – via le code génétique – qui forment une partition de la séquence d'ADN codante en trinucléotides (3-mots).

Ainsi, un modèle qui s'ajuste sur la fréquence des dinucléotides et la fréquence des codons a un sens biologique interprétable pour une séquence d'ADN codant un protéine.

Nous considérons une séquence $S = X_1X_2\cdots X_n$ $(n = 3n')$ découpée en codons, c'est à dire qu'à chaque base $X_i \in \mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, on associe la phase $i\%3$ (voir Chapitre 3). On définit donc le modèle, noté *M1_c* , par :

$$\mathrm{P}\{X_{3j} = c \mid X_{3j-2} = a,\ X_{3j-1} = b\} = \pi_3(ab, c) > 0,$$
$$\mathrm{P}\{X_{3j+k} = b \mid X_{3j+k-1} = a\} = \pi_k(a, b) > 0, \text{ pour } k = 1, 2.$$

Le modèle *M1_c* s'ajuste sur les comptages $N_{ab,1}$, $N_{ab,2}$ et $N_{abc,3}$, mais aussi sur les comptages $N_{ab,3}$ et N_{ab}. En effet,

$$N_{ab,3} = \sum_c N_{cab,3},$$

$$N_{ab} \quad = \quad N_{ab,1} + N_{ab,2} + N_{ab,3}\ ;$$

de la même façon, $N_{ab,2} = \sum_c N_{abc,3}$.
Ce modèle a 72 paramètres ; c'est un sous-modèle de *M2_3* (144 paramètres) et un sur-modèle de *M1_3* (36 paramètres).

On définit l'ensemble $\mathcal{C}$ des séquences de longueur n commençant par la première lettre de S, ayant des comptages de 2-mots identiques à ceux de S, $(N_{ab})_{a,b\in\mathcal{A}}$, et possédant le même nombres de codons $(abc, 3)$ que la séquence S : $N_{abc,3}$. On remarquera que toutes les séquences de $\mathcal{C}$ sont équiprobables, de probabilité

$$\mu(X_1) \prod_{a,b\in\mathcal{A},\, k=1,2} \pi_k(a,b)^{N_{ab,k}} \prod_{a,b,c\in\mathcal{A}} \pi_3(ab,c)^{N_{abc,3}}.$$

Nous rappelons que le nombre d'occurrences d'un mot W de longueur h dans S est défini par

$$N_W = \sum_{i=h}^{n} \mathbb{1}\{\text{un mot } W \text{ se termine en position } i \text{ dans } S\},$$

et que le nombre d'occurrences d'un mot W de longueur h en phase k dans S est

$$N_{W,k} = \sum_{i=h}^{n} \mathbb{1}\{\text{un mot } W \text{ se termine en position } i \text{ dans } S \text{ et } i = k[\text{modulo } 3]\}.$$

Nous allons étudier dans ce chapitre, la loi du nombre d'occurrences d'un mot W dans S conditionnellement à l'ensemble $\mathcal{C}$, $(N_W \,|\, \mathcal{C})$, ainsi que celle du comptage d'un mot W en phase k dans S, $(N_{W,k} \,|\, \mathcal{C})$. Pour cela, nous nous contenterons de calculer leurs espérances et variances conditionnelles pour construire les statistiques asymptotiquement gaussiennes centrées réduites (voir Chapitre 3) suivantes :

$$\frac{N_{W,k} - \mathbb{E}(N_{W,k} \,|\, \mathcal{C})}{\sqrt{\mathrm{Var}(N_{W,k} \,|\, \mathcal{C})}} \quad \text{et} \quad \frac{N_W - \mathbb{E}(N_W \,|\, \mathcal{C})}{\sqrt{\mathrm{Var}(N_W \,|\, \mathcal{C})}}.$$

Pour calculer l'espérance et la variance conditionnelles de (W, k), nous utilisons l'idée que Altshul et Erickson (1985) ont développé pour simuler des séquences de l'ensemble C. Ils ont remarqué qu'une fois que les lettres en phase 1 et 3 sont connues, il suffit, pour déterminer X_{3j+2}, de choisir uniformément parmi les codons de la forme $(X_{3j+1} \bullet X_{3j+3})$.
Ainsi, on considère l'ensemble $\mathcal{S}'$ des séquences distinctes obtenues en supprimant les lettres en phase 2 des séquences de $\mathcal{C}$. Chacune des séquences de $\mathcal{S}'$ commence ainsi par X_1, est de longueur $2n'$ et les comptages de leurs 2-mots sont fixés pour chacune des deux phases 1 et 3. Ces comptages sont les nombres $N_{ca,1}$ et $N_{a\bullet c,3}$ pour toutes les valeurs de a et c dans $\mathcal{A}$, où $N_{a\bullet c,3} = \sum_{b\in\mathcal{A}} N_{abc,3}$.

Dans la suite, le cardinal de l'ensemble $\mathcal{S}'$, noté $\#\mathcal{S}'$, va intervenir de nombreuses fois. C'est pourquoi, nous le calculons ici en appliquant la formule de Whittle (Théorème 2) comme nous l'avons fait en Section **3.3**. En effet, les séquences de $\mathcal{S}'$ sont des chaînes de Markov d'ordre 1 périodique de période 2 sur l'alphabet à huit lettres $\mathcal{A} \times \{1,3\}$, dont la matrice de transition $\overline{\Pi}'$ est

$$\overline{\Pi}' = \begin{pmatrix} 0 & \Pi'_3 \\ \Pi'_1 & 0 \end{pmatrix}$$

avec

$$\begin{aligned} \pi'_3(a,c) &= \sum_{b \in \mathcal{A}} \pi_2(a,b)\pi_3(ab,c) \quad \forall a,c \in \mathcal{A} , \\ \pi'_1(c,a) &= \pi_1(c,a) \quad \forall a,c \in \mathcal{A} . \end{aligned}$$

En remarquant que $N_{(a,1)+}$ défini par $\sum_c N_{a\bullet c,3}$ vaut exactement $N_{a,1}$, contrairement à $N_{(c,3)+} = \sum_a N_{ca,1}$, on obtient

$$\#\mathcal{S}' = \prod_{a \in \mathcal{A}} \frac{N_{a,1}!}{\prod_{c \in \mathcal{A}} N_{a\bullet c,3}!} \prod_{c \in \mathcal{A}} \frac{N_{(c,3)+}!}{\prod_{a \in \mathcal{A}} N_{ca,1}!} \times H_{(X_1,1)}(\mathcal{S}') \tag{4.1}$$

où $H_{(a,k)}(\mathcal{S}')$ est le cofacteur dans la matrice $I - \widehat{\overline{\Pi}'}$ de n'importe quel élément de la ligne (a,k) et $\widehat{\overline{\Pi}'}$ est la matrice de transition empirique associée à $\mathcal{S}'$:

$$\widehat{\overline{\Pi}'} = \begin{pmatrix} 0 & \left(\frac{N_{a\bullet c,3}}{N_{a,1}}\right)_{a,c \in \mathcal{A}} \\ \left(\frac{N_{ca,1}}{N_{(c,3)+}}\right)_{a,c \in \mathcal{A}} & 0 \end{pmatrix} .$$

Pour obtenir le cardinal de $\mathcal{C}$, il suffit de noter qu'à chaque séquence S' de $\mathcal{S}'$, correspond un nombre K de séquences de $\mathcal{C}$. Chacune de ces séquences est associée à une permutation distincte des codons de S qui conserve les sous-ensembles de codons dont la première et dernière lettre sont communes. On a donc

$$K = \prod_{a,c \in \mathcal{A}} \frac{N_{a\bullet c,3}!}{\prod_{b \in \mathcal{A}} N_{abc,3}!} \quad \text{et} \quad \#\mathcal{C} = K \times \#\mathcal{S}' .$$

Ainsi, dans un premier temps nous allons étudier les mots W en phase k dont la première et la dernière lettres ne sont pas en phase 2, c'est à dire tels que, ou bien la longueur h de W est multiple de trois et $k = 3$, ou bien $h = 2[\text{modulo } 3]$ et $k = 1$. Ensuite, nous généraliserons la méthode aux autres mots phasés (W,k).

4.2 Statistique d'écart au modèle pour deux types de mots (W, k) particuliers

Dans ce paragraphe, nous considérons les deux types de mots $\overline{W} = (W, k)$ suivant ($|W|$ représente la longueur du mot W) :

$$\begin{array}{lll} i) & \overline{W} = (W, 3) \text{ avec } h = |W| = 3\ell & \text{i.e. } W = C_1 \cdots C_\ell \text{ où } C_j = c_{j1}c_{j2}c_{j3} \\ ii) & \overline{W} = (W, 1) \text{ avec } h = |W| = 3\ell + 2 & \text{i.e. } W = u\, C_1 \cdots C_\ell\, v \text{ où } C_j = c_{j1}c_{j2}c_{j3}\ . \end{array}$$

Dans les deux cas, u et v représenteront respectivement la première et dernière lettre du mot W étudié.

Soit $\widetilde{W}$ le mot obtenu à partir de $\overline{W}$ en ignorant les lettres en phase 2:

$$\begin{array}{ll} i) & \widetilde{W} = c_{11} \bullet c_{13}\; c_{21} \bullet c_{23} \cdots c_{\ell 1} \bullet c_{\ell 3} \\ ii) & \widetilde{W} = u\; c_{11} \bullet c_{13}\; c_{21} \bullet c_{23} \cdots c_{\ell 1} \bullet c_{\ell 3}\; v\ . \end{array}$$

4.2.1 Calcul de $\mathbf{E}(N_{W,k} \,|\, \mathcal{C})$

On définit $p_{W,k}$ la probabilité d'observer $\overline{W}$ en i sachant qu'il y a déjà $\widetilde{W}$ en i, et conditionnellement à $\mathcal{C}$,

$$p_{W,k} = \mathrm{P}\{\text{un mot } \overline{W} \text{ se termine en } 3j + k \,|\, \mathcal{C},\ \text{un mot } \widetilde{W} \text{ se termine en } 3j + k\}\ .$$

Cette probabilité est donc celle d'observer les codons du mot W connaissant ses lettres en phases 1 et 3. Or, les séquences de $\mathcal{C}$ sont équiprobables et la distribution des lettres centrales des codons est indépendante d'un groupe $a \bullet c$, a et c fixés, à l'autre. Par conséquent, si $n_{abc,3}$ est le nombre de codons $(abc, 3)$ dans $\overline{W}$ et $n_{a\bullet c,3} = \sum_b n_{abc,3}$, on a

$$p_{W,k} = p_W = \prod_{a,c} \frac{\prod_b N_{abc,3}(N_{abc,3} - 1) \cdots (N_{abc,3} - n_{abc,3} + 1)}{N_{a\bullet c,3}(N_{a\bullet c,3} - 1) \cdots (N_{a\bullet c,3} - n_{a\bullet c,3} + 1)}\ , \tag{4.2}$$

et cette probabilité ne dépend pas de k.

Le calcul de $\mathbb{E}(N_{\overline{W}} \,|\, \mathcal{C})$ se déduit alors de

$$\begin{aligned} \mathbb{E}(N_{\overline{W}} \,|\, \mathcal{C}) &= \sum_{i=h}^{n} \mathrm{P}\{\text{un mot } \overline{W} \text{ se termine en } i \,|\, \mathcal{C}\} \\ &= p_W \sum_{i=h}^{n} \mathrm{P}\{\text{un mot } \widetilde{W} \text{ se termine en } i \,|\, \mathcal{C}\} \\ &= p_W \sum_{i=h}^{n} \mathrm{P}\{\text{un mot } \widetilde{W} \text{ se termine en } i \,|\, \mathcal{S}'\} \\ &= p_W\, \mathbb{E}(N_{\widetilde{W}} \,|\, \mathcal{S}')\ . \end{aligned} \tag{4.3}$$

L'espérance du second membre se calcule comme en **3.3.1** dans le cas d'un modèle 2-périodique, c'est à dire que l'on a :

$$\mathbb{E}(N_{\widetilde{W}} \mid \mathcal{S}') = N_{uv,k}(\mathcal{S}' \setminus \widetilde{W}) \, \frac{\#\mathcal{S}' \setminus \widetilde{W}}{\#\mathcal{S}'} \tag{4.4}$$

car $|\widetilde{W}|$ est pair ; si $|\widetilde{W}|$ est pair, il n'y a pas de décalage de phase dans les séquences de $\mathcal{S}' \setminus \widetilde{W}$. On rappelle que $\mathcal{S}' \setminus \widetilde{W}$ est l'ensemble des séquences distinctes obtenues en supprimant l'intérieur d'un mot $\widetilde{W}$ dans toutes les séquences de $\mathcal{S}'$ contenant au moins un mot $\widetilde{W}$; les comptages des dinucléotides phasés correspondant à ces séquences sont (suivant le type de $\overline{W}$) :

$$\begin{array}{rrcl} i) & N_{ca,1}(\mathcal{S}' \setminus \widetilde{W}) & = & N_{ca,1} - n_{ca,1} \\ & N_{a.c,3}(\mathcal{S}' \setminus \widetilde{W}) & = & N_{a.c,3} - n_{a.c,3} + \mathbb{1}\{(a,c) = (u,v)\} \\ ii) & N_{ca,1}(\mathcal{S}' \setminus \widetilde{W}) & = & N_{ca,1} - n_{ca,1} + \mathbb{1}\{(c,a) = (u,v)\} \\ & N_{a.c,3}(\mathcal{S}' \setminus \widetilde{W}) & = & N_{a.c,3} - n_{a.c,3} \end{array}$$

où $n_{ca,1}$ est le nombre de dinucléotide $(ca, 1)$ dans $\overline{W}$.
Dans (4.4), les cardinaux se calculent avec la formule de Whittle comme en (4.1).

Ainsi, en utilisant (4.3), (4.2) et (4.4), on a

$$\mathbb{E}(N_{\overline{W}} \mid \mathcal{C}) = \prod_{a,c} \frac{\prod_b N_{abc,3}(N_{abc,3} - 1) \cdots (N_{abc,3} - n_{abc,3} + 1)}{N_{a\bullet c,3}(N_{a\bullet c,3} - 1) \cdots (N_{a\bullet c,3} - n_{a\bullet c,3} + 1)} \, N_{uv,k}(\mathcal{S}' \setminus \widetilde{W}) \, \frac{\#\mathcal{S}' \setminus \widetilde{W}}{\#\mathcal{S}'} , \tag{4.5}$$

où k vaut 1 ou 3 suivant les cas $i)$ ou $ii)$.

4.2.2 Calcul de Var$(N_{W,k} \mid \mathcal{C})$

Calculons d'abord $\mathbb{E}(N^2_{\overline{W}} \mid \mathcal{C})$:

$$\mathbb{E}(N^2_{\overline{W}} \mid \mathcal{C}) = \sum_{i,i'} \mathrm{P}\{ \text{ un mot } \overline{W} \text{ se termine en } i \text{ et en } i' \mid \mathcal{C}\} .$$

On développe cette somme en trois termes T_1, T_2 et T_3 correspondant respectivement aux cas $i = i'$, $|i - i'| \geq h - 1$ et $|i - i'| < h - 1$; h étant la longueur de W. En reprenant les calculs développés dans **3.3.2**, on écrit directement

$$\begin{array}{rcl} T_1 & = & \mathbb{E}(N_{\overline{W}} \mid \mathcal{C}) , \\ T_3 & = & 2 \displaystyle\sum_{d=1}^{h-2} \delta(\overline{W}; d) \, \mathbb{E}(N_{\overline{W}^{(d)}\overline{W}} \mid \mathcal{C}) . \end{array}$$

On rappelle que $\delta(\overline{W}; d) = 1$ si le mot $\overline{W}$ peut apparaître à la fois en positions i et $i + d$ dans une séquence et dans ce cas on note $\overline{W}^{(d)}\overline{W}$ le mot obtenu en concaténant les d premières lettres de $\overline{W}$ avec $\overline{W}$, et sinon $\delta(\overline{W}; d) = 0$.
Si $\overline{W}$ est un mot de type $i)$ (resp. $ii)$) alors, si $\delta(\overline{W}; d) = 1$, $\overline{W}^{(d)}\overline{W}$ est lui aussi de type $i)$ (resp. $ii)$). Les termes T_1 et T_3 se calculent donc à l'aide de (4.5).

On définit $p_{W,W}$ la probabilité d'observer $\overline{W}$ simultanément aux positions i et i' sachant qu'il y a déjà $\widetilde{W}$ en i et i', et conditionnellement à $\mathcal{C}$,

$$\begin{aligned} p_{W,W} &= \mathrm{P}\{ \text{ un mot } \overline{W} \text{ se termine en } i \text{ et en } i' \,|\, \mathcal{C}, \text{ un mot } \widetilde{W} \text{ se termine en } i \text{ et en } i'\} \\ &= \prod_{a,c} \frac{\prod_b N_{abc,3}(N_{abc,3}-1)\cdots(N_{abc,3}-2n_{abc,3}+1)}{N_{a\bullet c,3}(N_{a\bullet c,3}-1)\cdots(N_{a\bullet c,3}-2n_{a\bullet c,3}+1)} \,. \end{aligned} \tag{4.6}$$

Le terme T_2 s'écrit alors

$$\begin{aligned} T_2 &= \sum_{|i-i'|\geq h-1} \mathrm{P}\{\text{un mot } \overline{W} \text{ se termine en } i \text{ et en } i' \,|\, \mathcal{C}\} \\ &= p_{W,W} \sum_{|i-i'|\geq h-1} \mathrm{P}\{\text{un mot } \widetilde{W} \text{ se termine en } i \text{ et en } i' \,|\, \mathcal{C}\} \\ &= p_{W,W} \sum_{|i-i'|\geq h-1} \mathrm{P}\{\text{un mot } \widetilde{W} \text{ se termine en } i \text{ et en } i' \,|\, \mathcal{S}'\} \\ &= p_{W,W} \,.\, N_{uv,k}(\mathcal{S}' \setminus \widetilde{W} \setminus \widetilde{W})(N_{uv,k}(\mathcal{S}' \setminus \widetilde{W} \setminus \widetilde{W}) - 1) \,\frac{\#\mathcal{S}' \setminus \widetilde{W} \setminus \widetilde{W}}{\#\mathcal{S}'} \end{aligned}$$

La quatrième égalité est démontrée en Section **3.3.2**. Pour calculer $\#\mathcal{S}'\setminus\widetilde{W}\setminus\widetilde{W}$, il suffit d'utiliser le méthode de calcul de $\#\mathcal{S}' \setminus \widetilde{W}$.

Ainsi,

$$\begin{aligned} \mathbb{E}(N_{\overline{W}}^2 \,|\, \mathcal{C}) &= \mathbb{E}(N_{\overline{W}} \,|\, \mathcal{C}) \;+\; 2\sum_{d=1}^{h-2} \delta(\overline{W}; d)\mathbb{E}(N_{\overline{W}^{(d)}\overline{W}} \,|\, \mathcal{C}) \\ &\quad + \; p_{W,W} \,.\, N_{uv,k}(\mathcal{S}' \setminus \widetilde{W} \setminus \widetilde{W})(N_{uv,k}(\mathcal{S}' \setminus \widetilde{W} \setminus \widetilde{W}) - 1)\frac{\#\mathcal{S}' \setminus \widetilde{W} \setminus \widetilde{W}}{\#\mathcal{S}'} \end{aligned} \tag{4.7}$$

avec

$$\begin{array}{lll} i) & h = 3\ell & \text{et } k = 3, \\ ii) & h = 3\ell + 2 & \text{et } k = 1. \end{array}$$

D'où

$$\mathrm{Var}(N_{W,k} \,|\, \mathcal{C}) = \mathbb{E}(N_{\overline{W}}^2 \,|\, \mathcal{C}) - \mathbb{E}(N_{\overline{W}} \,|\, \mathcal{C})^2 \,.$$

4.3 Statistique d'écart au modèle : généralisation

4.3.1 Pour un mot W en phase k

On veut, de la même façon, calculer l'espérance et la variance conditionnelle à $\mathcal{C}$ de $N_{\overline{W}}$ où $\overline{W} = (W,k)$ est quelconque ; on note $h = |W|$. La méthode qui va être utilisée consiste à se ramener aux cas traités en **4.2**, c'est à dire à exprimer $N_{\overline{W}}$ comme combinaison linéaire de comptages de mots de types $i)$ ou $ii)$. Pour cela, il faut étudier les mots $\overline{W}$ séparément, suivant leur longueur modulo 3 et leur phase :

• $h = 3\ell + 2$

$$\begin{array}{rll}
\textbf{(I)} & k=1 & : \overline{W} \text{ est un mot de type } ii) \\
\textbf{(II)} & k=2 & : N_{\overline{W}} = \sum_a N_{Wa,3} \text{ et } (Wa,3) \text{ est un mot de type } i) \\
\textbf{(III)} & k=3 & : N_{\overline{W}} = \sum_a N_{aW,3} \text{ et } (aW,3) \text{ est un mot de type } i)
\end{array}$$

• $h = 3\ell + 1$

$$\begin{array}{rll}
\textbf{(IV)} & k=1 & : N_{\overline{W}} = \sum_{a,b} N_{Wab,3} \text{ et } (Wab,3) \text{ est un mot de type } i) \\
\textbf{(V)} & k=2 & : N_{\overline{W}} = \sum_{a,b} N_{aWb,3} \text{ et } (aWb,3) \text{ est un mot de type } i) \\
\textbf{(VI)} & k=3 & : N_{\overline{W}} = \sum_{a,b} N_{abW,3} \text{ et } (abW,3) \text{ est un mot de type } i)
\end{array}$$

• $h = 3\ell$

$$\begin{array}{rll}
\textbf{(VII)} & k=1 & : N_{\overline{W}} = \sum_{a,b,c} N_{aWbc,3} \text{ et } (aWbc,3) \text{ est un mot de type } i) \\
\textbf{(VIII)} & k=2 & : N_{\overline{W}} = \sum_{a,b,c} N_{abWc,3} \text{ et } (abWc,3) \text{ est un mot de type } i) \\
\textbf{(IX)} & k=3 & : \overline{W} \text{ est un mot de type } i)
\end{array}$$

Dans chacun des cas ci-dessus, on s'aperçoit qu'il suffit d'utiliser la formule (4.5) pour calculer l'espérance conditionnelle de $\overline{W}$; quant à la variance conditionnelle, on a besoin, soit de la formule (4.7) (cas **I** et **IX**) , soit de la proposition suivante qui nous permet de calculer la covariance conditionnelle de deux mots $\overline{W}$ et $\overline{W'}$ de type $i)$ (cas **II**, **III**, **IV**, **V**, **VI**, **VII** et **VIII**).

Proposition 1 *Soient $\overline{W} = (W,k)$ et $\overline{W'} = (W',k')$ deux mots de type i) ou ii) avec $|W| = h$ et $|W'| = h'$. On note u et v (resp. u' et v') la première et dernière lettre de W (resp. W') et on pose $\overline{uv} = (uv,k)$, $\overline{u'v'} = (u'v',k')$.*
Alors, on a

$$\mathbb{E}(N_{\overline{W}}N_{\overline{W'}}\,|\,\mathcal{C}) = \sum_{d=2-h'}^{h-2} \delta(\overline{W};\overline{W'};d)\; \mathbb{E}(N_{\overline{W}^{(d)}\overline{W'}}\,|\,\mathcal{C})$$

$$+ \quad p_{W,W'}\,.\,N_{\overline{uv}}(\mathcal{S}'\setminus\widetilde{W}\setminus\widetilde{W'})(N_{\overline{u'v'}}(\mathcal{S}'\setminus\widetilde{W}\setminus\widetilde{W'}) - \delta(\overline{uv};\overline{u'v'};0))\;\frac{\#\mathcal{S}'\setminus\widetilde{W}\setminus\widetilde{W'}}{\#\mathcal{S}'}$$

où

$$p_{W,W'} = \prod_{a,c} \frac{\prod_b N_{abc,3}(N_{abc,3}-1)\cdots(N_{abc,3}-n_{abc,3}-n'_{abc,3}+1)}{N_{a\bullet c,3}(N_{a\bullet c,3}-1)\cdots(N_{a\bullet c,3}-n_{a\bullet c,3}-n'_{a\bullet c,3}+1)} \tag{4.8}$$

et $n'_{abc,3}$ est le nombre de codons $(abc,3)$ du mot $\overline{W'}$ et $n'_{a\bullet c,3} = \sum_b n'_{abc,3}$.

Preuve : Par analogie aux notations précédentes, $\delta(\overline{W};\overline{W'};d) = 1$ si un mot $\overline{W}$ et un mot $\overline{W'}$ peuvent apparaître simultanément aux positions i et $i+d$ dans une séquence ; dans ce cas, on note $\overline{W}^{(d)}\overline{W'}$ le mot concaténé (il est de type *i)* ou *ii)*). Dans le cas contraire, $\delta(\overline{W};\overline{W'};d) = 0$. Par définition,

$$\mathbb{E}(N_{\overline{W}}N_{\overline{W'}}\,|\,\mathcal{C}) = \sum_{i,i'} \mathrm{P}\{\text{un mot } \overline{W} \text{ se termine en } i \text{ et un mot } \overline{W'} \text{ se termine en } i'\,|\,\mathcal{C}\}\,.$$

On va suivre le schéma utilisé pour calculer $\mathbb{E}(N^2_{\overline{W}}\,|\,\mathcal{C})$ (formule (4.7)) à savoir écrire la somme ci-dessus sous la forme $T'_1 + T'_2$, où T'_1 correspond au cas où $\overline{W}$ et $\overline{W'}$ se chevauchent sur plus d'une lettre, et T'_2, au cas où $\overline{W}$ et $\overline{W'}$ apparaissent de façon non chevauchante dans la séquence (le chevauchement d'une lettre est toutefois autorisé).

On obtient ainsi

$$T'_1 = \sum_{d=2-h'}^{h-2} \delta(\overline{W};\overline{W'};d)\; \mathbb{E}(N_{\overline{W}^{(d)}\overline{W'}}\,|\,\mathcal{C})\,,$$

et en notant $p_{W,W'}$ la probabilité d'observer $\overline{W}$ en i et $\overline{W'}$ en i' sachant qu'il y a $\widetilde{W}$ en i et $\widetilde{W'}$ en i', et conditionnellement à $\mathcal{C}$, on a

$$\begin{aligned} T'_2 &= p_{W,W'}\,.\sum_{\substack{i-i'\geq h-1\\ \text{ou}\\ i-i'\leq 1-h'}} \mathrm{P}\{\text{ un mot } \widetilde{W} \text{ termine en } i \text{ et un mot } \widetilde{W'} \text{ termine en } i'\,|\,\mathcal{S}'\} \\ &= p_{W,W'}\,.\,N_{\overline{uv}}(\mathcal{S}'\setminus\widetilde{W}\setminus\widetilde{W'})(N_{\overline{u'v'}}(\mathcal{S}'\setminus\widetilde{W}\setminus\widetilde{W'}) - \delta(\overline{uv};\overline{u'v'};0))\;\frac{\#\mathcal{S}'\setminus\widetilde{W}\setminus\widetilde{W'}}{\#\mathcal{S}'}\;; \end{aligned}$$

cette dernière égalité résulte d'un calcul en **3.3.2**.
Ceci démontre la proposition. □

4.3.2 Pour un mot W

On veut maintenant calculer l'espérance et la variance conditionnelles à $\mathcal{C}$ de N_W, où W est de longueur h. On va utiliser la décomposition suivante

$$N_W = N_{W,1} + N_{W,2} + N_{W,3} \,,$$

puis utiliser, suivant la longueur h de W, celles données au paragraphe précédent, c'est à dire

- $h = 3\ell + 2$

$$N_W = N_{W,1} + \sum_a N_{Wa,3} + \sum_a N_{aW,3} \,,$$

- $h = 3\ell + 1$

$$N_W = \sum_{a,b} N_{Wab,3} + \sum_{a,b} N_{aWb,3} + \sum_{a,b} N_{abW,3} \,,$$

- $h = 3\ell$

$$N_W = \sum_{a,b,c} N_{aWbc,3} + \sum_{a,b,c} N_{abWc,3} + N_{W,3} \,,$$

de façon à se ramener à des mots de type $i)$ ou $ii)$. Le calcul de $\mathbb{E}(N_W \mid \mathcal{C})$ est ainsi immédiat avec la formule (4.5) ; le calcul de $\mathrm{Var}(N_W \mid \mathcal{C})$ s'obtient avec la proposition 1.

Deuxième partie

Approximation de Poisson composée du comptage d'un mot dans une séquence d'ADN

Chapitre 5

Approximations de Poisson et combinatoire des mots

5.1 Introduction

Dans ce chapitre, nous supposons que l'espérance du nombre d'occurrences d'un mot W est bornée lorsque l'on fait tendre la longueur de la séquence vers l'infini. Le mot W est donc considéré comme un événement rare dans la séquence, ce qui nous permet d'étudier des approximations de Poisson.

La distribution du nombre d'occurrences $N(W)$ d'un mot W, ainsi que celle du nombre de renouvellements de W, aussi appelé nombre d'occurrences non chevauchantes de W, ont été très étudiées dans le cas d'une suite $(X_i)_{i=1\cdots n}$ où les v.a. X_i sont indépendantes et équidistribuées sur un certain alphabet (Godbole (1991), Chryssaphinou et Papastavridis (1988a), Chryssaphinou et Papastavridis (1988b), Hirano et Aki (1993), Godbole et Schaffner (1993), Fu (1993)). La convergence vers une loi de Poisson a été souvent établie dans deux situations asymptotiques différentes, soit lorsque les probabilités de chacune des lettres de l'alphabet tend vers 0, soit quand la longueur du mot étudié tend vers l'infini. Ces travaux utilisent soit les fonctions génératrices, soit la méthode de Chen-Stein qui permet d'approcher une somme de variables aléatoires de Bernoulli non nécessairement indépendantes par une v.a. de Poisson, tout en donnant un contrôle de l'erreur d'approximation. Cette méthode est en fait une extension, développée par Chen (1975), de la méthode de Stein (1972) qui donne une majoration explicite de l'erreur commise , en terme de distance en variation totale, lorsqu'une somme de v.a. dépendantes est approchée par une loi Normale. Un grand nombre d'applications de cette méthode peut être trouvé dans Arratia *et al.* (1990b).

Sous l'hypothèse de dépendance markovienne de la suite $(X_i)_{i=1\cdots n}$, la plupart des auteurs ont étudié la distribution du nombre d'occurrences du motif $\mathbb{1}_h = 11\cdots1$ (h fois) dans une chaîne de Markov d'ordre 1 à deux états $\{0,1\}$ (Godbole (1991), Godbole et Schaffner (1993),

Rajarshi (1974), Hirano et Aki (1993)). En notant les probabilités de transition $\alpha = \pi(1,1)$ et $\beta = \pi(0,1)$, Godbole (1991) montre que, sous l'hypothèse naturelle d'événement rare $n\beta\alpha^{h-1} \to \lambda$, la loi de $N(\mathbb{1}_h)$ tend vers la loi de Poisson de paramètre λ, pourvu que α et β tendent vers 0 lorsque n tend vers l'infini ; il utilise la méthode de Chen-Stein. Au lieu d'imposer des conditions sur les paramètres α et β de la chaîne de Markov, Rajarshi (1974) montre, en utilisant les fonctions génératrices, que si $n \to +\infty$ et $h = h_n \to +\infty$ tels que $n\alpha^{h-1} \to \lambda$ alors $N(\mathbb{1}_h)$ a asymptotiquement une loi de Poisson de paramètre $\dfrac{\lambda\alpha\beta}{2-\alpha-\beta}$. Le cadre asymptotique de ce dernier travail correspond à celui de notre étude en ce sens que les paramètres de la chaîne de Markov d'ordre m que l'on ajuste sur la séquence d'ADN sont fixés. Aussi, pour assurer la condition $\mathbb{E}N(W) \to \lambda$, nous devons considérer des mots W_n dont la longueur h_n tend vers l'infini avec n.
Dans ce cadre, Geske *et al.* (1995) considèrent une chaîne de Markov d'ordre 1 sur un espace d'état fini quelconque, et obtiennent une approximation de la loi de $N(W)$ par une loi de Poisson composée, mais uniquement pour une certaine catégorie de mots que l'on précisera à la fin de ce chapitre. Nous reprenons ici ce même problème, mais en considérant tous les mots possibles au lieu de nous restreindre à certains mots "simples". Pour obtenir ce résultat, ils appliquent la version processus de la méthode de Chen-Stein telle qu'elle est présentée dans Arratia *et al.* (1989). Une méthode alternative est celle proposée par Barbour *et al.* (1992b) ; ces derniers prolongent la méthode de Stein pour évaluer directement l'erreur d'approximation d'un processus, généré par des v.a. de Bernoulli, par un processus de Poisson.

Plus généralement, nous étudions ici la distribution du nombre d'occurrences $N(W)$ d'un mot W quelconque dans une chaîne de Markov homogène d'ordre m sur un alphabet fini, en l'occurrence $\mathcal{A}$. Puisque l'on peut considérer une chaîne de Markov d'ordre m sur l'espace d'état $\mathcal{A}$ comme une chaîne de Markov d'ordre 1 sur l'espace d'état $\mathcal{A}^m$, nous traitons en détail le cas $m = 1$, puis nous montrons comment se généralisent les résultats pour un ordre m quelconque. De façon progressive, nous démontrons que l'approximation de la loi de $N(W)$ par une loi de Poisson composée est bien meilleure que par une simple loi de Poisson, surtout si le mot W possède une structure très chevauchante. En effet, si la structure du mot W est telle que deux occurrences de W peuvent se chevaucher dans la séquence, il est clair que les occurrences des mots W dans des "paquets", que l'on appellera "trains" de mots W, ont une probabilité différente des occurrences isolées ; le nombre de ces trains est en fait approximativement poissonnien. Ce phénomène est mentionné par plusieurs auteurs (Arratia *et al.* (1990a), Arratia *et al.* (1990b), Godbole et Schaffner (1993), Karlin et Ost (1987)).

Ce chapitre se décompose de la manière suivante. Tout d'abord, nous étudions le lien entre les deux comptages suivants : le nombre d'occurrences $N(W)$ défini comme $\sum_{i\in I} Y_i$ où $I = \{1,\ldots,n-h+1\}$ et Y_i est la v.a. qui vaut 1 si une occurrence de W commence en position i dans la séquence et 0 sinon, et le nombre de trains $\widetilde{N}^*(W)$ de W défini, aux erreurs de bord près, comme $\sum_{i\in I} \widetilde{Y}_i$ où $\widetilde{Y}_i = Y_i(1-Y_{i-1})\cdots(1-Y_{i-h+1})$. Dans un second temps, nous appliquons la méthode de Chen-Stein pour approcher le processus des occurrences de trains de W, $(\widetilde{Y}_i)_{i\in I}$, par un processus de Poisson $(Z_i)_{i\in I}$ convenable. Ensuite, nous distinguons les trains de W suivant le nombre k d'occurrences de W dont ils sont composés, ce qui revient à attribuer à chaque train un type $k \in \{1, 2, \cdots\}$; on définit ainsi les v.a. $\widetilde{Y}_i^{(k)}$ qui valent 1 si un train de type k commence en position i dans la séquence et 0 sinon, et on montre que le processus $(\widetilde{Y}_i^{(k)})_{i,k}$ ainsi formé est approché par un processus de Poisson $(Z_i^{(k)})_{i,k}$ approprié. Enfin, il suffit de remarquer que $N(W)$ est peu différent, en un sens que l'on précisera, de $\sum_{i,k} k\,\widetilde{Y}_i^{(k)}$ dont la distribution est approchée par celle de $\sum_{i,k} k\,Z_i^{(k)}$, c'est à dire une loi de Poisson composée. Il est clair que l'introduction des variables aléatoires $\widetilde{Y}_i$ et $\widetilde{Y}_i^{(k)}$ implique, tout au long de ce travail, l'utilisation importante de résultats combinatoires sur les mots, décrivant en particulier les différentes possibilités d'un mot à se chevaucher lui-même.

Il serait intéressant de comparer nos résultats avec ceux que l'on obtiendrait en utilisant les travaux de Barbour *et al.* (1992a) et Roos (1994). Ils étendent la méthode de Stein pour majorer la distance en variation totale entre une somme de Bernoulli et une loi de Poisson composée ayant un nombre fini de composantes. Dans le cas de mots très simples, l'erreur d'approximation est du même ordre que celle que nous obtenons, mais les composantes de leur loi de Poisson composée sont différentes des nôtres et ne correspondent pas aux k-trains.

5.2 Comptages d'un mot dans une séquence

Soit $S = X_1X_2\cdots X_n$ une séquence d'ADN représentée par la succession de ses bases X_i à valeurs dans l'alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. On s'intéresse aux occurrences, dans cette séquence, d'un mot $W = w_1\cdots w_h$ de longueur h, notée aussi $|W|$, sur l'alphabet $\mathcal{A}$. Suivant que l'on compte toutes les occurrences de W dans la séquence, ou seulement certaines, on obtient des comptages différents. C'est pourquoi, nous décrivons maintenant trois types de comptages.

5.2.1 Description de trois types de comptages

• Le comptage naturel

Le comptage naturel, noté $N(W)$, tient compte de toutes les occurrences, même chevauchantes, du mot W dans la séquence.

Définition 1

$$N(W) = \sum_{i=1}^{n-h+1} Y_i$$

où Y_i vaut 1 si un mot W commence en position i dans la séquence et 0 sinon, c'est à dire $Y_i = \mathbb{1}\{X_iX_{i+1}\cdots X_{i+h-1} = W\}$.

Exemple : Si $S =$ TGAACAAACAACAAT, $N(\mathtt{AACAA}) = 3$ (positions 3, 7, 10).

• Le comptage sans chevauchement

Le comptage sans chevauchement, noté $\overset{o}{N}(W)$, est le nombre maximum d'occurrences non chevauchantes de W dans la séquence. Autrement dit, on compte de gauche à droite les occurrences des *renouvellements* de W, où un renouvellement de W est défini comme étant une occurrence de W qui ne chevauche pas un précédent renouvellement de W. (*Renewal theory de Feller, 1968*).

Définition 2

$$\overset{o}{N}(W) = \sum_{i=1}^{n-h+1} R_i$$

où R_i est défini séquentiellement par

$$R_i = \begin{cases} 1 & si\ X_i X_{i+1} \cdots X_{i+h-1} = W \ et\ R_{i-k} = 0 \ pour\ k = 1, \ldots, h-1 \\ 0 & sinon, \end{cases}$$

c'est à dire par $R_i = Y_i \prod_{k=1}^{h-1} (1 - R_{i-k})$ *avec la convention* $R_i = 0$ *pour* $i \in \{2 - h, \ldots, 0\}$.

À partir du moment où le mot W peut se chevaucher lui-même, le calcul de R_i dépend de tous les X_j, $j < i + h$.

Exemple Si $S =$`TGAACAAACAACAAT`, $\overset{o}{N}$(`AACAA`) $= 2$ (positions 3, 10).

• Le comptage des trains

Ce comptage, noté $\widetilde{N}^*(W)$, ne tient compte que de certaines occurrences non chevauchantes du mot W ; en effet, un mot W en position i dans la séquence est compté si et seulement si il n'y a pas de mots W dans les $h - 1$ positions précédentes. Autrement dit, un mot W en position i est compté si et seulement si $Y_i(1 - Y_{i-1}) \cdots (1 - Y_{i-h+1}) = 1$ lorsque $i \geq h$ ou $Y_i(1 - Y_{i-1}) \cdots (1 - Y_1) = 1$ lorsque $i < h$.

Définition 3

$$\widetilde{N}^*(W) = \sum_{i=1}^{n-h+1} \widetilde{Y}_i^*$$

où $\widetilde{Y}_i^* = Y_i \prod_{k=1}^{h-1} (1 - Y_{i-k})$ *si* $i \geq h$ *et* $\widetilde{Y}_i^* = Y_i \prod_{k=1}^{i-1} (1 - Y_{i-k})$ *si* $i < h$.

Exemple Si $S =$`TGAACAAACAACAAT`, $\widetilde{N}^*$(`AACAA`) $= 1$ (position 3).

Sur cet exemple, on voit bien que parmi les trois occurrences du mot `AACAA` dans S, seule la première est comptée, car la troisième chevauche la deuxième, qui elle-même chevauche la première. On dit ainsi qu'on observe un train de mots W en position i si un motif C, composé d'occurrences chevauchantes de W, apparaît en i et ne chevauche de mots W, ni devant, ni derrière. On appelle *tête de train* la première occurrence d'un train. Sur l'exemple ci-dessus, l'occurrence de `AACAA` en position 3 est la tête du train `AACAAACAACAA`. Ainsi, $\widetilde{Y}_i^* = 1$ si et seulement si il y a une occurrence de W en position i et celle-ci est une tête de train.

Remarques

– On a l'inégalité

$$\widetilde{Y}_i^* \leq R_i \leq Y_i\,, \quad i = 1, \ldots, n-h+1,$$

qui est une égalité si le mot W ne peut pas se chevaucher.

– De plus,

$$\begin{array}{rll} Y_i & \text{dépend de} & X_i, \ldots, X_{i+h-1}\,, \\ R_i & \text{dépend de} & X_1, X_2, \ldots, X_i, \ldots, X_{i+h-1}\,, \\ \widetilde{Y}_i^* & \text{dépend de} & X_{i-h+1}, \ldots, X_i, \ldots, X_{i+h-1}. \end{array}$$

Pour qu'un mot W soit un renouvellement, il faut tenir compte non seulement de sa position dans le train auquel il appartient, mais aussi de la structure de ce train ; ainsi R_i ne présente pas de structure markovienne par rapport à la filtration $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$, contrairement à $\widetilde{Y}_i^*$ et Y_i. Dans la suite, nous ne parlerons plus du nombre de renouvellements.

5.2.2 Lien entre $N(W)$ et $\widetilde{N}^*(W)$

L'objectif est ici d'obtenir une formule, aussi simple que possible, reliant les deux comptages $N(W)$ et $\widetilde{N}^*(W)$. On pourrait développer $\widetilde{Y}_i^*$ en fonction des $Y_i, Y_{i-1}, \ldots, Y_{i-h+1}$, mais il apparaît 2^{h-1} termes de la forme $Y_i Y_{i-j_1} \cdots Y_{i-j_\ell}$ dont, généralement, beaucoup sont nuls. Leur nullité est liée aux différents chevauchements impossibles de W avec lui-même. En étudiant précisément les chevauchements possibles, nous allons simplifier l'expression $Y_i \prod_{k=1}^{h-1}(1 - Y_{i-k})$, en diminuant le nombre de termes.

Dans ce paragraphe, nous supposons que les bases X_{-h+2}, …, X_{-1}, X_0 de la séquence sont observées, de façon à considérer les variables

$$\widetilde{Y}_i = Y_i \prod_{k=1}^{h-1}(1 - Y_{i-k})\,, \quad i \in I.$$

Les problèmes de bords sont résolus ultérieurement.

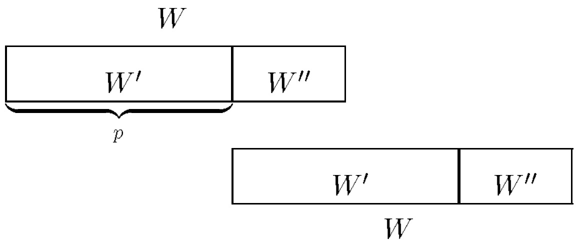
Un mot W est chevauchant si il peut exister deux occurrences de ce mot aux positions i et j avec $|j-i| < h$; on définit ci-dessous la notion de *période* pour caractériser la distance $|j-i|$ entre ces deux occurrences.

Définition 4 *Soit $W = w_1 \cdots w_h$ un mot de longueur h ; $p \in \{1, \ldots, h-1\}$ est dite* période[1] *de W si et seulement si $w_i = w_{i+p}$, $i \in \{1, \ldots, h-p\}$.*
À chaque période p de W, on associe le mot $W^{(p)} = w_1 \cdots w_p$ que l'on appellera racine *de W.*

Soit $\mathcal{P}(W)$ l'ensemble des périodes de W ; ainsi $\mathcal{P}(W) = \emptyset$ si et seulement si le mot W ne se chevauche pas.

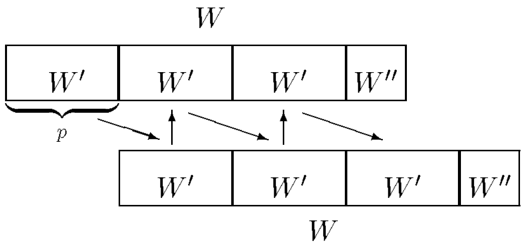
Nous mettons en évidence, ci-dessous, la structure périodique d'un mot chevauchant. Soit $p \in \mathcal{P}(W)$; on distingue deux cas :

• si $p > \frac{h}{2}$, on est dans la situation du schéma suivant :



et le mot W se décompose ainsi : $W = W'W''$ où W'' est un début de W', c'est à dire que $W' = W''W'''$; en changeant de notations, $W' = yz$ et $W = yzy$.

• si $p \leq \frac{h}{2}$, on est dans la situation du schéma suivant :



et le mot W se décompose ainsi : $W = (W')^r W''$ avec $r \geq 2$ et W'' est le début de W' (W'' peut être le mot vide) ; avec les notations précédentes , on a $W' = yz$ et $W = (yz)^r y$.

On a donc obtenu le résultat suivant :

Proposition 2 *Pour toute période $p \in \mathcal{P}(W)$, W se décompose de la manière suivante :*

$$W = (yz)^r y \text{ où } \begin{cases} |yz| = p \\ r \geq 1 \\ |z| \geq 1 \end{cases}$$

La décomposition associée à la plus petite période de W est appelée décomposition canonique *de W.*

1. Dans le vocabulaire usuel de la Théorie combinatoire des mots, la longueur d'un mot est considérée comme une période de ce mot. Cependant, pour simplifier nos énoncés, nous ne considérons que les périodes différentes de la longueur du mot.

L'ensemble des périodes d'un mot W est généralement donné de façon récursive (Guibas et Odlyzko, 1981a). Dès que l'on connaît la période minimale p_0 et la racine minimale $|yz| = W^{(p_0)}$ de W, le théorème ci-dessous précise comment obtenir les autres périodes possibles à partir des périodes du mot yzy plus court.

Théorème 10 *Soit p_0 la période minimale de W et $(yz)^r y$ sa décomposition canonique ; on a donc $|yz| = p_0$. L'ensemble des périodes de W est la réunion des trois ensembles disjoints suivants :*

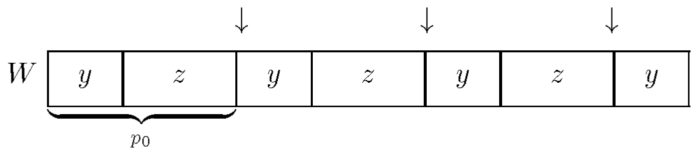
(i) $\{jp_0,\ 1 \leq j \leq r\}$

(ii) $\{rp_0 + p(y),\ p(y) \in \mathcal{P}(y)\}$

(iii) *si* $r > 1$, $\{(r-1)p_0 + p(yzy),\ p(yzy) \in \mathcal{P}(yzy)$ *et* $|y| < p(yzy) < |yz|\}$.

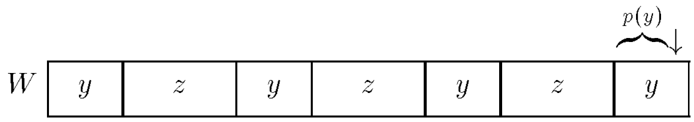
Nous illustrons ces trois cas par des schémas.

- Cas (i) : jp_0 est période de W $(1 \leq j \leq r)$; $(yz)^j$ est ainsi racine de W :
  (Les flèches indiquent les positions où un nouveau mot W peut commencer).



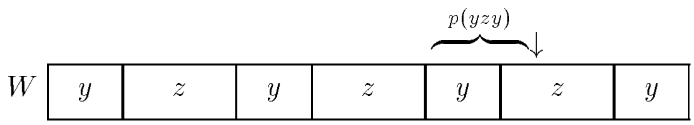
  Exemple : si $W =$ ACG ACG ACG AC, la période minimale est 3 ; 6 et 9 sont aussi des périodes de W ; les racines associées sont ACG, ACGACG et ACGACGACG.

- Cas (ii) : si le suffixe y se chevauche, avec une période $p(y)$, alors $rp_0 + p(y)$ est période de W. Par conséquent, pour toute racine y' de y ($y = y'y''$ et $|y'| \in \mathcal{P}(y)$), le mot $(yz)^r y'$ est racine de W :



  Exemple : si $W =$ ACAT ACAT ACA, la période minimale est 4 ; 10 est aussi période de W associée à la racine ACATACATAC

- Cas (iii) : si $r \geq 2$ et si le suffixe yzy a une période $p(yzy)$ telle que $|y| < p(yzy) < |yz|$, alors $(r-1)p_0 + p(yzy)$ est période de W. Donc, pour toute racine de yzy de la forme yz' $(1 \leq |z'| < |z|)$, le mot $(yz)^{r-1}yz'$ est racine de W :

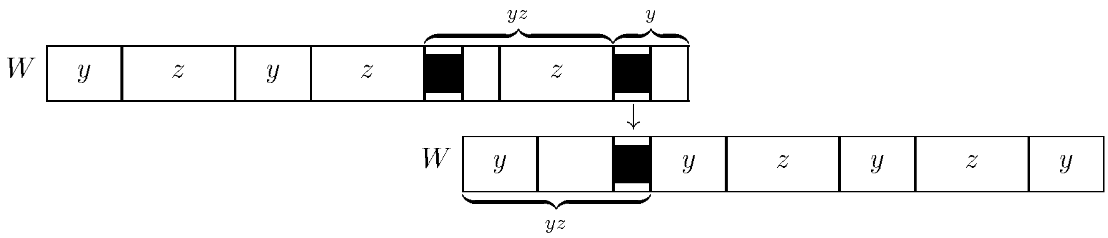


Exemple : si W = ATAATATAATA, la période minimale est 5 ; 8 est aussi période de W associée à la racine ATAATATA.

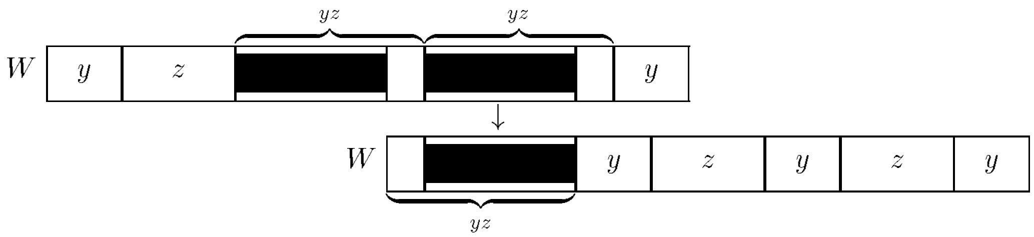
Si yz' est racine de yzy, cela n'implique pas que z' soit racine de z : en effet, sur cet exemple $y =$ A, $z =$ TAAT et $z' =$ TA n'est pas une racine de z.

Pour démontrer le Théorème 2, il suffit de démontrer que les autres chevauchements, c'est à dire d'autres périodes pour W, sont impossibles. Une démonstration formelle est donnée dans Gélis (communication personnelle, 1994), mais il n'est pas difficile de le voir sur un schéma. En fait, si l'on considère la décomposition canonique $W = (yz)^r y$, les autres chevauchements sont de deux sortes ;

– soit le second mot W commence dans un des y (sauf le dernier) :



– soit il commence dans un des z (sauf le dernier) :



Dans les deux situations représentées ci-dessus, yz se décompose sous la forme

$$yz = xx' = x'x$$

où les mots x et x' sont non vides, ce qui implique, en utilisant le théorème 1.3.2 de Lothaire (1983), l'existence d'un autre mot x_0 et d'un entier $\ell > 1$ tel que $yz = (x_0)^\ell$; ceci contredit le fait que yz soit la racine minimale de W. $\square$

Nous énonçons ci-dessous deux résultats fondamentaux pour la suite, qui seront utilisés notamment pour démontrer les Propositions 3 et 4.

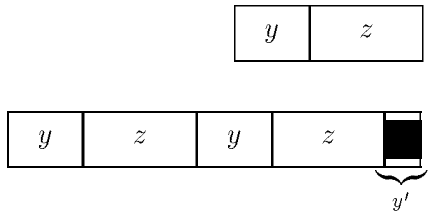
Définition 5 *Une période qui n'est pas multiple de la période minimale est dite* principale. *On note $\mathcal{P}'(W)$ l'ensemble des périodes principales.*

Résultat 1 Pour qu'un mot W soit tête de train dans une séquence, il suffit d'exclure juste avant lui dans la séquence les *racines principales* associées aux périodes principales de W.

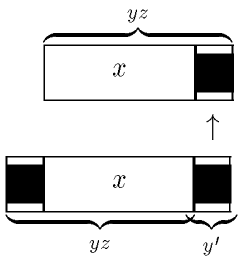
Résultat 2 L'occurrence simultanée de deux racines principales différentes , $W^{(p)}$ et $W^{(q)}$ où $p, q \in \mathcal{P}'(W)$, devant une même occurrence de W dans la séquence est impossible.

Le Résultat 1 est trivial. On vérifie ci-dessous le Résultat 2, dans le cas où $p = p_0$ et q est de type (ii), puis dans le cas où p et q sont de type (ii). Pour les autres cas, la preuve se fait de la même façon.

- Supposons que l'on ait $W^{(p_0)} = yz$ en position $i - p_0$ et $W^{(q)} = (yz)^r y'$ en position $i - q$ avec y' racine de y :

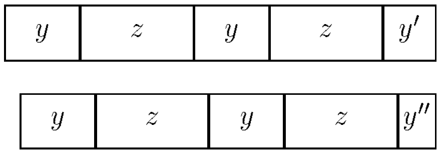
  

  Puisque y' est une racine de y, on a

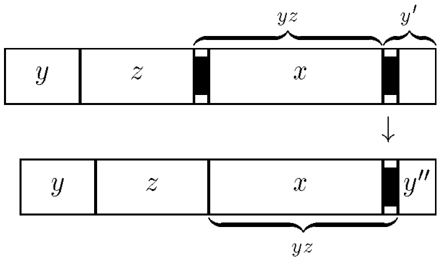
  

  et yz se décompose ainsi sous la forme $yz = xy' = y'x$ où x est un mot non vide ; en utilisant à nouveau le théorème 1.3.2 de Lothaire (1983) on en déduit que yz n'est pas la racine minimale de W, d'où la contradiction.

- Supposons que l'on ait $W^{(p)} = (yz)^r y'$ en position $i-p$ et $W^{(q)} = (yz)^r y''$ en position $i-q$ avec y' et y'' deux racines différentes de y :

  On a, par conséquent, le schéma suivant

  et on conclut à nouveau en remarquant que yz s'écrit $yz = x'x = xx'$. □

Proposition 3 *Soit $W = w_1 \cdots w_h$ un mot de longueur h et $\mathcal{P}'(W)$ l'ensemble de ses périodes principales. Pour tout $i \in I = \{1, \ldots, n-h+1\}$, on note*

$$Y_i(W) = \mathbb{1}\{X_i X_{i+1} \cdots X_{i+h-1} = W\} \text{ et } \widetilde{Y}_i(W) = Y_i(W) \prod_{k=1}^{h-1} (1 - Y_{i-k}(W)).$$

Alors, on a

$$\widetilde{Y}_i(W) = Y_i(W) - \sum_{p \in \mathcal{P}'(W)} Y_{i-p}(W^{(p)}W).$$

où $W^{(p)}W$ est le motif composé $w_1 \cdots w_p w_1 \cdots w_h$.

Preuve: Par le Résultat 1, pour qu'un mot W en position i soit tête de train, il faut et il suffit qu'il y ait W en i et qu'il n'y ait aucun $W^{(p)}$ en $i-p$ avec $p \in \mathcal{P}'(W)$. Ceci nous donne

$$\begin{aligned} \widetilde{Y}_i(W) &= Y_i(W) \mathbb{1} \left\{ \bigcap_{p \in \mathcal{P}'(W)} \{\text{il n'y a pas de mot } W^{(p)} \text{ en } i-p\} \right\} \\ &= Y_i(W) \left(1 - \mathbb{1} \left\{ \bigcup_{p \in \mathcal{P}'(W)} \{\text{il y a un mot } W^{(p)} \text{ en } i-p\} \right\} \right) \end{aligned}$$

Or, d'après le Résultat 2, on a une union d'événements disjoints, d'où

$$\widetilde{Y}_i(W) = Y_i(W) \left(1 - \sum_{p \in \mathcal{P}'(W)} Y_{i-p}(W^{(p)}) \right).$$

□

Remarque : On vient ainsi de démontrer l'égalité

$$Y_i(W)(1 - Y_{i-1}(W)) \cdots (1 - Y_{i-h+1}(W)) = Y_i(W) - \sum_{p \in \mathcal{P}'(W)} Y_{i-p}(W^{(p)}W)$$

ce qui revient à développer le membre de gauche et à éliminer les termes nuls correspondant à des chevauchements simultanés impossibles. Le nombre de termes est ainsi passé de 2^{h-1} à au plus $1 + [h/2]$. La formule ci-dessus était difficile à trouver en regardant directement les produits $Y_i Y_{i-j_1} \cdots Y_{i-j_\ell}$.

Bien que le nombre de trains $\widetilde{N}^*(W)$ soit défini par $\sum_{i \in I} \widetilde{Y}_i^*(W)$, où $\widetilde{Y}_i^*$ diffère quelque peu de $\widetilde{Y}_i$ lorsque $1 \leq i < h$, la Proposition 3 donne tout de même la relation

$$\widetilde{N}^*(W) = N(W) - \sum_{p \in \mathcal{P}'(W)} N(W^{(p)}W)$$

où $N(\cdot)$ est le nombre d'occurrences dans la séquence.

5.3 Approximations de Poisson pour les comptages

5.3.1 Méthode de Chen-Stein

Les deux théorèmes ci-dessous, sont basés sur la méthode d'approximation de Chen-Stein, due à Stein (1972) et Chen (1975), qui permet d'obtenir un majorant de l'erreur d'approximation lorsque l'on utilise une loi de Poisson pour approcher une somme de v.a. de Bernoulli non indépendantes. Cette erreur d'approximation est majorée en terme de distance en variation des deux distributions. La distance en variation entre les lois $\mathcal{L}(Y)$ et $\mathcal{L}(Z)$ de deux variables aléatoires, à valeurs entières positives, Y et Z est définie par

$$\begin{aligned} d_{\mathrm{VT}}(\mathcal{L}(Y),\mathcal{L}(Z)) &= \sup_{||f||=1} |\mathbb{E}f(Y)-\mathbb{E}f(Z)| \\ &= 2\sup_{A\subset\mathbb{N}} |\mathrm{P}\{Y\in A\}-\mathrm{P}\{Z\in A\}| \\ &= 2\inf \mathrm{P}^*\{Y\neq Z\}, \end{aligned}$$

où la borne inf est prise pour l'ensemble des lois P^* dont les marginales sont les lois de Y et Z.

Notations : Soit I un ensemble d'indices fini ou dénombrable ; pour chaque $i\in I$, on considère Y_i une variable aléatoire de Bernoulli de paramètre $p_i = \mathrm{P}\{Y_i=1\} = 1-\mathrm{P}\{Y_i=0\} > 0$ et $B_i\subset I$ un voisinage de i contenant i. On note

$$\begin{aligned} b_1 &= \sum_{i\in I}\sum_{j\in B_i} p_i p_j \\ b_2 &= \sum_{i\in I}\sum_{j\in B_i\backslash\{i\}} \mathbb{E}(Y_iY_j) \\ b_3' &= \sum_{i\in I}\mathbb{E}\left|\mathbb{E}\left(Y_i-p_i\mid \sum_{j\in B_i^c} Y_j\right)\right| \\ b_3 &= \sum_{i\in I}\mathbb{E}|\mathbb{E}(Y_i-p_i|\sigma(Y_j, j\in B_i^c))|. \end{aligned}$$

Théorème 11 (Arratia *et al.* (1989)) *Soient $N=\sum_{i\in I} Y_i$ et Z une variable aléatoire de Poisson d'espérance $\lambda=\mathbb{E}(N)\in]0,+\infty[$. On a*

$$\begin{aligned} d_{VT}(\mathcal{L}(N),\mathcal{L}(Z)) &\leq 2\left[(b_1+b_2)\frac{1-e^{-\lambda}}{\lambda} + b_3'(1\wedge 1.4\lambda^{-1/2})\right] \\ &\leq 2(b_1+b_2+b_3). \end{aligned}$$

Le théorème suivant est la version processus du théorème ci-dessus.

Théorème 12 (Arratia *et al.* (1990b)) *Pour tout $i \in I$, on considère les v.a. Z_i indépendantes et distribuées suivant une loi de Poisson de paramètre p_i. La distance en variation entre le processus de Bernoulli dépendant $\mathbb{Y} = (Y_i)_{i \in I}$ et le processus de Poisson $\mathbb{Z} = (Z_i)_{i \in I}$ d'intensité $p_{(.)}$ vérifie*

$$d_{VT}(\mathcal{L}(\mathbb{Y}), \mathcal{L}(\mathbb{Z})) \leq 2(2b_1 + 2b_2 + b_3).$$

5.3.2 Approximation de la loi de $\widetilde{N}^*(W)$ par une loi de Poisson (M1)

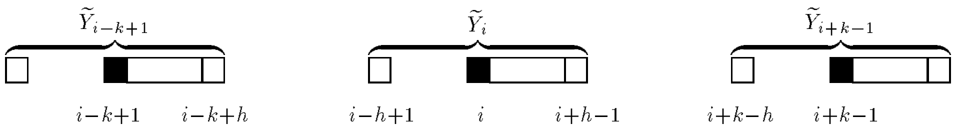
On suppose que la séquence $S = X_1 X_2 \dots X_n$ est extraite d'une chaîne de Markov infinie $\{X_i\}_{i=-\infty}^{i=+\infty}$ stationnaire d'ordre 1 (modèle M1), dont on observe aussi $X_{-h+2} \dots X_{-1} X_0$. Soit $W = w_1 \cdots w_h$ un mot de longueur h ; on notera $u = w_1$ et $v = w_h$.

Nous démontrons ici que, si la séquence est assez longue et la longueur h du mot W suffisamment grande, la loi du nombre $\widetilde{N}^*(W)$ de trains de W peut être approchée par une loi de Poisson. Par contre, une approximation de Poisson pour la distribution de $N(W)$ n'est pas bonne, surtout si le mot W est très chevauchant car il aura tendance à former des trains ; une approximation par une loi de Poisson composée est meilleure.

Nous appliquons en parallèle la méthode de Chen-Stein pour les deux variables $N(W) = \sum_{i \in I} Y_i$ et $\widetilde{N}(W) = \sum_{i \in I} \widetilde{Y}_i$, où $I = \{1, \dots, n-h+1\}$ et $\widetilde{Y}_i = Y_i(1 - Y_{i-1}) \cdots (1 - Y_{i-h+1})$. On note $\mu(W) = \mathbb{E}(Y_i)$ et $\widetilde{\mu}(W) = \mathbb{E}(\widetilde{Y}_i)$. La Proposition 3 fournit l'expression de $\widetilde{\mu}(W)$ suivante

$$\widetilde{\mu}(W) = \mu(W) - \sum_{p \in \mathcal{P}'(W)} \mu(W^{(p)} W). \tag{5.1}$$

Choix du voisinage Pour tout $i \in I$, on définit le voisinage $B_i = \{j \in I, |j - i| < k\}$ où $k \geq 2h - 1$, de sorte que l'ensemble d'indices ℓ des X_ℓ intervenant dans $\widetilde{Y}_i$ et celui des indices intervenant dans $\widetilde{Y}_j$ soient disjoints si $j \notin B_i$, autrement dit, $j - h + 1 > i + h - 1$ ou $i - h + 1 > j + h - 1$ pour $j \notin B_i$ (voir schéma).



Calcul de b_1

$$\begin{aligned} b_1(Y) &= \sum_{i \in I} \sum_{j \in B_i} \mathbb{E}(Y_i)\mathbb{E}(Y_j) = \sum_{i \in I} \sum_{j \in B_i} \mu^2(W) \\ &\leq (n - h + 1)(2k - 1)\mu^2(W) \end{aligned}$$

car $\text{card}(I) = n - h + 1$ et $\text{card}(B_i) \leq 2k - 1$, $\forall i$ (on a l'égalité lorsque $k \leq i \leq n - k + 1$).

De la même façon,

$$\begin{aligned} b_1(\widetilde{Y}) &= \sum_{i\in I}\sum_{j\in B_i}\mathbb{E}(\widetilde{Y}_i)\mathbb{E}(\widetilde{Y}_j) = \sum_{i\in I}\sum_{j\in B_i}\widetilde{\mu}^2(W) \\ &\leq (n-h+1)(2k-1)\widetilde{\mu}^2(W). \end{aligned}$$

Par conséquent, quel que soit le comptage considéré, on a

$$b_1 \leq (n-h+1)(2k-1)\mu^2(W).$$

Calcul de b_2

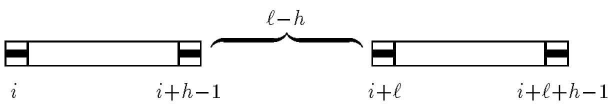
$$b_2(Y) = \sum_{i\in I}\sum_{j\in B_i\setminus\{i\}}\mathbb{E}(Y_iY_j)\,.$$

On fait apparaître les termes Y_iY_j où $|i-j| < h$ qui correspondent à un chevauchement des occurrences de W en i et j (si il est possible), et les termes Y_iY_j où $|i-j| \geq h$ qui correspondent à deux occurrences disjointes du mot W.

$$\begin{aligned} b_2(Y) &= \sum_{i\in I}\left(\sum_{j\in\{i-h+1,\ldots,i+h-1\}\setminus\{i\}}\mathbb{E}(Y_iY_j) + \sum_{j\in B_i\setminus\{i-h+1,\ldots,i+h-1\}}\mathbb{E}(Y_iY_j)\right) \\ &= \sum_{i\in I}\left(2\sum_{p\in\mathcal{P}(W)}\mu(W^{(p)}W) + 2\sum_{\ell=h}^{k-1}\mathbb{E}(Y_iY_{i+\ell})\right), \end{aligned}$$

or

$$\begin{aligned} \mathbb{E}(Y_iY_{i+\ell}) &= \mathrm{P}\{\text{un mot } W \text{ commence en } i \text{ et un autre en } i+\ell\} \\ &= \mu(W)\frac{\Pi^{\ell-h+1}(v,u)}{\mu(u)}\mu(W)\,,\ \text{si } \ell \geq h \end{aligned}$$



d'où

$$\begin{aligned} b_2(Y) &= \sum_{i\in I}\left(2\sum_{p\in\mathcal{P}(W)}\mu(W^{(p)}W) + 2\sum_{\ell=h}^{k-1}\mu^2(W)\frac{\Pi^{\ell-h+1}(v,u)}{\mu(u)}\right) \\ &= 2(n-h+1)\left(\sum_{p\in\mathcal{P}(W)}\mu(W^{(p)}W) + \sum_{\ell=h}^{k-1}\mu^2(W)\frac{\Pi^{\ell-h+1}(v,u)}{\mu(u)}\right). \end{aligned}$$

On s'aperçoit ici que, dès que $n\mu(W) = \mathrm{O}(1)$, l'approximation de la loi de $N(W)$ par une loi de Poisson sera mauvaise à cause du (ou des) terme(s) en $n\mu(W^{(p)}W)$ dans l'expression de

$b_2(Y)$ car, dès lors que p peut prendre des petites valeurs (si W est très chevauchant), on a $b_2(Y) = \mathrm{O}(n\mu(W)) = \mathrm{O}(1)$.

Calculons maintenant $b_2(\widetilde{Y})$.

$$b_2(\widetilde{Y}) = \sum_{i\in I} \sum_{j\in B_i\setminus\{i\}} \mathbb{E}(\widetilde{Y}_i\widetilde{Y}_j) \ ;$$

par définition de $\widetilde{Y}_i$ et $\widetilde{Y}_j$, on a $\widetilde{Y}_i\widetilde{Y}_j = 0$ dès que $|i-j| < h$, d'où

$$\begin{aligned}
b_2(\widetilde{Y}) &= \sum_{i\in I} \sum_{j\in B_i\setminus\{i-h+1,\ldots,i+h-1\}} \mathbb{E}(\widetilde{Y}_i\widetilde{Y}_j) \\
&= 2\sum_{i\in I}\sum_{\ell=h}^{k-1} \mathbb{E}(\widetilde{Y}_i\widetilde{Y}_{i+\ell}) \\
&\leq 2.\mathrm{card}(I).\frac{\mu^2(W)}{\mu(u)} \sum_{\ell=1}^{k-h} \Pi^{\ell}(v,u) \\
&\leq \frac{2}{\mu(u)}(n-h+1)(k-h)\mu^2(W).
\end{aligned}$$

Calcul de $b_3(Y)$

Si on note $\mathcal{Y}_i = \sigma(Y_j,\ j\in B_i^c)$ et $\mathcal{X}_i = \sigma(X_1,\ldots,X_{i-k+h-1},X_{i+k},\ldots,X_n)$, on a par définition

$$b_3(Y) = \sum_{i\in I} \mathbb{E}|\ \mathbb{E}(Y_i - \mathbb{E}(Y_i)|\mathcal{Y}_i)\ |. \tag{5.2}$$

En remarquant que $\mathcal{Y}_i \subset \mathcal{X}_i$, on se ramène à un conditionnement par rapport à $\mathcal{X}_i$ en utilisant les propriétés de l'espérance conditionnelle

$$\mathbb{E}|\ \mathbb{E}(Y_i - \mathbb{E}(Y_i)|\mathcal{Y}_i)\ | \leq \mathbb{E}|\ \mathbb{E}(Y_i - \mathbb{E}(Y_i)|\mathcal{X}_i)\ |$$

puis on applique la propriété de Markov :

$$\mathbb{E}|\ \mathbb{E}(Y_i - \mathbb{E}(Y_i)|\mathcal{Y}_i)\ | \leq \mathbb{E}|\ \mathbb{E}(Y_i|X_{i-k+h-1},X_{i+k}) - \mathbb{E}(Y_i)\ |\,.$$

Pour calculer le membre de droite, on développe l'espérance suivant les valeurs de $X_{i-k+h-1}$ et X_{i+k} puis on calcule les probabilités apparaissant à l'aide de la matrice de transition de la chaîne de Markov :

$\mathbb{E}|\ \mathbb{E}(Y_i|X_{i-k+h-1},X_{i+k}) - \mathbb{E}(Y_i)\ | =$

$$\begin{aligned}
&= \sum_{x,y} |\ \mathbb{E}(Y_i|X_{i-k+h-1}=x,X_{i+k}=y) - \mu(W)\ | \times \mathrm{P}\{X_{i-k+h-1}=x, X_{i+k}=y\} \\
&= \sum_{x,y} \left|\mu(x)\Pi^{k-h+1}(x,u)\frac{\mu(W)}{\mu(u)}\Pi^{k-h+1}(v,y) - \mu(W)\mu(x)\Pi^{2k-h+1}(x,y)\right| \\
&= \mu(W)\sum_{x,y}\mu(x)\left|\frac{1}{\mu(u)}\Pi^{k-h+1}(x,u)\Pi^{k-h+1}(v,y) - \Pi^{2k-h+1}(x,y)\right|.
\end{aligned}$$

En diagonalisant la matrice de transition $\Pi = PDP^{-1}$ avec D diagonale, on a

$$\Pi^h = PD^hP^{-1} = \sum_{t=1}^{4} \alpha_t^h P I_t P^{-1} = \sum_{t=1}^{4} \alpha_t^h Q_t$$

avec les valeurs propres $\alpha_1 = 1$ et $|\alpha_t| < 1$ pour $t \neq 1$ d'où

$$\begin{aligned}
\mathbb{E}| \quad & \mathbb{E}(Y_i \,|\, X_{i-k+h-1}, X_{i+k}) - \mathbb{E}(Y_i) \,| = \\
&= \mu(W) \sum_{x,y} \mu(x) \left| \frac{1}{\mu(u)} \sum_{t=1}^{4} \alpha_t^{k-h+1} Q_t(x,u) \sum_{t'=1}^{4} \alpha_{t'}^{k-h+1} Q_{t'}(v,y) - \sum_{t=1}^{4} \alpha_t^{2k-h+1} Q_t(x,y) \right| \\
&= \mu(W) \sum_{x,y} \mu(x) \left| \frac{1}{\mu(u)} \sum_{(t,t')\neq(1,1)} \alpha_t^{k-h+1} \alpha_{t'}^{k-h+1} Q_t(x,u) Q_{t'}(v,y) - \sum_{t=2}^{4} \alpha_t^{2k-h+1} Q_t(x,y) \right|
\end{aligned} \tag{5.3}$$

car $\forall a, b \in \mathcal{A},\ Q_1(a,b) = \mu(b)$ et pour $t = t' = 1$ les termes correspondant s'annulent.

Ainsi, en notant α la seconde plus grande valeur propre en module ($|\alpha| < 1$), on obtient à partir de (5.2)

$$b_3(Y) \leq C(u,v,h)\ (n-h+1)\mu(W)|\alpha|^{k-h+1}$$

avec

$$C(u,v,h) = \sum_{x,y} \mu(x) \left| \frac{1}{\mu(u)} \sum_{(t,t')\neq(1,1)} \frac{\alpha_t^{k-h+1}\alpha_{t'}^{k-h+1}}{\alpha^{k-h+1}} Q_t(x,u) Q_{t'}(v,y) - \sum_{t=2}^{4} \frac{\alpha_t^{2k-h+1}}{\alpha^{k-h+1}} Q_t(x,y) \right|.$$

Calcul de $b_3(\widetilde{Y})$

Pour calculer $b_3(\widetilde{Y})$, il suffit de remarquer que, si on définit l'ensemble $\mathcal{U}$ des mots de longueur h qui ne se terminent pas par une des racines principales de W, autrement dit

$$\mathcal{U} = \{u_1 \cdots u_h \,|\, \forall p \in \mathcal{P}'(W),\ u_{h-p+1} \cdots u_h \neq W^{(p)}\}, \tag{5.4}$$

la variable $\widetilde{Y}_i$ s'écrit sous la forme

$$\widetilde{Y}_i = \sum_{U \in \mathcal{U}} Y_{i-h}(UW). \tag{5.5}$$

Ainsi, en reprenant les techniques utilisées pour le calcul de $b_3(Y)$, on obtient directement la majoration

$$b_3(\widetilde{Y}) \leq \sum_{i \in I} \mathbb{E} \left| \sum_{U \in \mathcal{U}} \mathbb{E}(Y_{i-h}(UW) \mid X_{i-k+h-1}, X_{i+k-h+1}) - \mathbb{E}(Y_{i-h}(UW)) \right|,$$

puis en utilisant la formule (5.3), on a

$$\begin{aligned} b_3(\widetilde{Y}) \;\leq\; & \sum_{i\in I}\sum_{x,y}\mu(x)\sum_{U\in\mathcal{U}}\mu(UW)\left|\frac{1}{\mu(u_1)}\sum_{(t,t')\neq(1,1)}\alpha_t^{k-2h+1}\alpha_{t'}^{k-2h+1}Q_t(x,u_1)Q_{t'}(v,y)\right. \\ & \left. -\sum_{t=2}^{4}\alpha_t^{2k-h+2}Q_t(x,y)\right|. \end{aligned}$$

Si α est la seconde valeur propre la plus grande en module, de la matrice de transition Π, on peut majorer l'expression entre les valeurs absolues par $C(x,y)|\alpha|^{k-2h+1}$, où $C(x,y)$ est une constante. Par conséquent, la majoration ci-dessus se réduit à

$$b_3(\widetilde{Y}) \leq |\alpha|^{k-2h+1}\sum_{i\in I}\sum_{x,y}\mu(x)C(x,y)\sum_{U\in\mathcal{U}}\mu(UW).$$

Or, $\widetilde{\mu} = \sum_{U\in\mathcal{U}}\mu(UW)$ d'après la formule (5.5), d'où la majoration de $b_3(\widetilde{Y})$

$$b_3(\widetilde{Y}) \leq C(n-h+1)\widetilde{\mu}(W)|\alpha|^{k-2h+1} \tag{5.6}$$

où C est une constante.

Conclusion

Si on reprend les différentes majorations de b_1, b_2 et b_3, on constate les choses suivantes :

– que ce soit $b_1(Y)$ ou $b_1(\widetilde{Y})$, on a $b_1 \leq (n-h+1)(2k-1)\mu^2(W)$, donc en se plaçant dans le cadre asymptotique $n\mu(W) = \mathrm{O}(1)$, $h = \mathrm{o}(n)$ et $k = \mathrm{o}(n)$, $b_1(Y)$ et $b_1(\widetilde{Y})$ tendent vers zéro.

– De plus, on a vu que si $n\mu(W) = \mathrm{O}(1)$ alors $b_2(Y)$ ne tendait pas nécessairement vers zéro ; par contre, dans le cadre asymptotique précisé ci-dessus, $b_2(\widetilde{Y})$ tend vers zéro car $b_2(\widetilde{Y}) \leq \mathrm{O}(n(k-h)\mu^2(W))$.

– Enfin, si $n\mu(W) = \mathrm{O}(1)$ alors $b_3(Y)$ (resp. $b_3(\widetilde{Y})$) tend vers zéro pourvu que $k-h+1$ (resp. $k-2h+1$) tende vers $+\infty$ et $h = \mathrm{o}(n)$.

En utilisant le Théorème 11 pour la variable $\widetilde{N}(W) = \sum_{i=1}^{n-h+1}\widetilde{Y}_i$, puis en remarquant que les comptages $\widetilde{N}(W)$ et $\widetilde{N}^*(W)$ ont la même distribution asymptotique car

$$d_{\mathrm{VT}}(\mathcal{L}(\widetilde{N}(W)), \mathcal{L}(\widetilde{N}^*(W))) = 2\mathrm{P}\{\widetilde{N}(W) \neq \widetilde{N}^*(W)\} \leq 2h\,\mu(W),$$

on établit le résultat suivant :

Théorème 13 *Soit W un mot de longueur h. On a*

$$\begin{aligned} & d_{VT}(\mathcal{L}(\widetilde{N}^*(W)), \mathcal{P}o(n\widetilde{\mu}(W))) \;\leq \\ & \qquad (n-h+1)\left(B_1k\widetilde{\mu}^2(W) + B_2(k-h)\mu^2(W) + B_3\widetilde{\mu}(W)|\alpha|^{k-2h+1}\right) + 2h\mu(W). \end{aligned}$$

avec $k \geq 2h-1$ et les quantités B_1, B_2, B_3 sont des O(1).

Remarque : le Théorème 13 nous permettrait de calculer une majoration de l'erreur à distance finie, mais en pratique, la majoration des constantes n'est pas calculée avec finesse. Par conséquent, nous nous placerons dans le cadre asymptotique du corollaire suivant, obtenu en prenant $k = 3h$.

Corollaire 1 *On considère une suite de mots W_n de longueur h_n telle que $n\mu(W_n) = O(1)$ et $h_n = o(n)$. On a*

$$\lim_{n \to +\infty} d_{VT}(\mathcal{L}(\widetilde{N}^*(W_n)), \mathcal{P}o(n\widetilde{\mu}(W_n))) = 0$$

Remarque : La condition $n\mu(W_n) = O(1)$ impose que h_n tende vers l'infini à une vitesse au moins égale à $\log n$.

5.3.3 Approximation de la loi de $N(W)$ par une loi de Poisson composée (M1)

Les techniques utilisées au paragraphe précédent permettent, en utilisant le Théorème 12, d'approcher le processus $\widetilde{\mathbb{Y}} = (\widetilde{Y}_i)_{i \in I}$ des occurrences des trains par un processus de Poisson $\mathbb{Z} = (Z_i)_{i \in I}$ d'intensité $\widetilde{\mu} = \mathbb{E}(\widetilde{Y}_i)$.

L'idée consiste maintenant à colorer chacun des trains de la suite $\{X_i\}_{i=-\infty}^{i=+\infty}$ suivant le nombre d'occurrences du mot W dont ils sont composés, ce qui nous permettra de reconstituer le nombre $N(W)$ d'occurrences de W observées dans la séquence $S = X_1 \cdots X_n$. Ainsi, pour tout $i \in I$ et tout $k \geq 1$, on définit les variables

$$\widetilde{Y}_i^{(k)} = \mathbb{1}\{ \text{ un train composé d'exactement } k \text{ mots } W \text{ commence en } i\}.$$

De cette façon,

$$\widetilde{Y}_i = \sum_{k \geq 1} \widetilde{Y}_i^{(k)}$$

et $\sum_{i \in I} \widetilde{Y}_i^{(k)}$ est le nombre $\widetilde{N}^{(k)}$ de trains composés d'exactement k occurrences du mot W et qui commencent dans $X_1 \cdots X_{n-h+1}$.

Soit $\mathbb{Z}^* = (Z_i^{(k)})_{i \in I, k \geq 1}$ le processus de Poisson tel que les $Z_i^{(k)}$ soient des v.a de Poisson indépendantes de paramètre $\widetilde{\mu}_k = \mathbb{E}(\widetilde{Y}_i^{(k)})$. Le Théorème 12 nous permet d'approcher le processus $\widetilde{\mathbb{Y}}^* = (\widetilde{Y}_i^{(k)})_{i \in I, k \geq 1}$ par le processus de Poisson $\mathbb{Z}^*$ avec un contrôle de l'erreur d'approximation :

$$d_{\mathrm{VT}}(\mathcal{L}(\widetilde{\mathbb{Y}}^*), \mathcal{L}(\mathbb{Z}^*)) \leq 2(2b_1^* + 2b_2^* + b_3^*) \ ;$$

les quantités b_1^*, b_2^* et b_3^* sont calculées et majorées un peu plus loin.

Par ailleurs, les variables aléatoires $Z^{(k)}$ définies par

$$Z^{(k)} = \sum_{i \in I} Z_i^{(k)}$$

sont indépendantes et de loi de Poisson d'espérance $\widetilde{\lambda}_k = (n-h+1)\widetilde{\mu}_k$ et par définition $f(\mathbb{Z}^*) = \sum_{k\geq 1} k\, Z^{(k)}$ suit la loi de Poisson composée $\mathrm{CP}(\Lambda)$ où $\Lambda \equiv \sum_{k\geq 1} \widetilde{\lambda}_k \delta_k$ est une mesure sur $\{1, 2, \ldots\}$.
Par conséquent, si on définit $N^*(W) \stackrel{\text{déf}}{=} \sum_{k\geq 1} k\, \widetilde{N}^{(k)}$, c'est à dire $N^*(W) = f(\widetilde{\mathbb{Y}}^*)$, on peut approcher la loi de $N^*(W)$ par la loi de Poisson composée $\mathrm{CP}(\Lambda)$:

$$
\begin{aligned}
d_{\mathrm{VT}}(\mathcal{L}(N^*(W)), \mathrm{CP}(\Lambda)) &= d_{\mathrm{VT}}(\mathcal{L}(f(\widetilde{\mathbb{Y}}^*)), \mathcal{L}(f(\mathbb{Z}^*))) \\
&\leq d_{\mathrm{VT}}(\mathcal{L}(\widetilde{\mathbb{Y}}^*), \mathcal{L}(\mathbb{Z}^*)) \\
&\leq 2(2b_1^* + 2b_2^* + b_3^*).
\end{aligned}
$$

On a introduit la nouvelle notation $N^*(W)$ car le comptage $N^*(W)$ n'est pas égal au comptage naturel $N(W)$. En effet, on peut avoir la situation suivante : un train commence avant X_{n-h+2} et se termine au delà de X_n dans la suite infinie $\{X_i\}_{i=-\infty}^{i=+\infty}$. Ce train censuré, s'il existe, est unique et, parmi toutes les occurrences de W qui le composent et qui sont comptées dans $N^*(W)$, seules celles observées dans $X_1 \cdots X_n$ sont comptées dans $N(W)$. Dans ce cas, il y a nécessairement une occurrence de W dans une des positions $n-2h+3, \ldots, n-h+1$ de la séquence $X_1X_2\cdots X_n$. L'autre cas introduisant une différence entre $N^*(W)$ et $N(W)$ est celui où un train commence avant X_1 et se termine au delà de X_{h-1}, et dans ce cas, il y a nécessairement une occurrence de W dans les $h-1$ premières positions de la séquence $X_1X_2\cdots X_n$. Par conséquent, on majore

$$
\mathrm{P}\{N(W) \neq N^*(W)\} \leq 2h\, \mu(W).
$$

Cependant, on verra qu'en moyenne, ces deux biais se compensent car $N(W)$ et $N^*(W)$ ont la même espérance.

En utilisant l'inégalité triangulaire, on obtient

$$
d_{\mathrm{VT}}(\mathcal{L}(N(W)), \mathrm{CP}(\Lambda)) \leq 2(2b_1^* + 2b_2^* + b_3^*) \; + 4h\, \mu(W). \tag{5.7}
$$

Avant de montrer que les quantités b_1^*, b_2^* et b_3^* tendent vers zéro, nous explicitons la loi de Poisson composée limite ci-dessus.

Calcul de $\widetilde{\mu}_k$

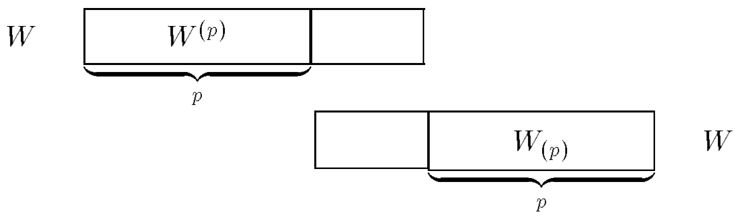
Afin de caractériser la loi de Poisson composée $\mathrm{CP}(\Lambda)$, on calcule $\widetilde{\mu}_k = \mathbb{E}(\widetilde{Y}_i^{(k)})$, pour tout $k \geq 1$. Pour cela, on utilise l'expression de $\widetilde{Y}_i^{(k)}$ donnée par la proposition ci-dessous, dont la Proposition 3 devient une conséquence.

Proposition 4 *Soient un mot W et $\mathcal{P}'(W)$ l'ensemble de ses périodes principales. Soit $\mathcal{C}_k$ l'ensemble de tous les motifs composés d'une succession de k occurrences chevauchantes de W. Alors, on a*

$$\begin{aligned}\widetilde{Y}_i^{(k)}(W) &= \sum_{C\in\mathcal{C}_k}\Bigg(Y_i(C) - \sum_{p\in\mathcal{P}'(W)} Y_{i-p}\left(W^{(p)}\,C\right) - \sum_{q\in\mathcal{P}'(W)} Y_i\left(C\,W_{(q)}\right) \\ &\qquad + \sum_{p,q\in\mathcal{P}'(W)} Y_{i-p}\left(W^{(p)}\,C\,W_{(q)}\right)\Bigg),\end{aligned}$$

où $W_{(q)}$ représente les q dernières lettres de W.

Preuve : Pour qu'un train, d'exactement k occurrences de W, commence en position i, il faut et il suffit qu'un motif $C \in \mathcal{C}_k$ apparaisse en i, qu'il ne soit précédé par aucune racine principale $W^{(p)}$, et suivi par aucun suffixe $W_{(q)}$ où q est une période principale de W. En effet, lorsque l'on considère un chevauchement du type



on s'aperçoit que l'occurrence de droite est précédée par $W^{(p)}$ et que celle de gauche est suivie par $W_{(p)}$. Or, on a vu que deux chevauchements principaux distincts (c'est à dire correspondant à deux périodes principales différentes) ne peuvent se produire simultanément sur une même occurrence de W. De ce fait, deux motifs distincts C et C' de $\mathcal{C}_k$ ne peuvent apparaître à la fois en position i, ce qui implique

$$\begin{aligned}\widetilde{Y}_i^{(k)}(W) &= \sum_{C\in\mathcal{C}_k} Y_i(C)\,\mathbb{1}\left\{\left\{\bigcap_{p\in\mathcal{P}'(W)}\{W^{(p)} \text{ en } i-p\}^{\mathrm{c}}\right\}\bigcap\left\{\bigcap_{q\in\mathcal{P}'(W)}\{W_{(q)} \text{ en } i+|C|\}^{\mathrm{c}}\right\}\right\} \\ &= \sum_{C\in\mathcal{C}_k} Y_i(C)\left(1-\mathbb{1}\left\{\left\{\bigcup_{p\in\mathcal{P}'(W)}\{W^{(p)} \text{ en } i-p\}\right\}\bigcup\left\{\bigcup_{q\in\mathcal{P}'(W)}\{W_{(q)} \text{ en } i+|C|\}\right\}\right\}\right) \\ &= \sum_{C\in\mathcal{C}_k}\left(Y_i(C)-\mathbb{1}\left\{\left\{\bigcup_{p\in\mathcal{P}'(W)}\{W^{(p)}C \text{ en } i-p\}\right\}\bigcup\left\{\bigcup_{q\in\mathcal{P}'(W)}\{CW_{(q)} \text{ en } i\}\right\}\right\}\right).\end{aligned}$$

Pour tout $p, q \in \mathcal{P}'(W)$, les événements $\{W^{(p)}C \text{ en } i-p\}$ et $\{CW_{(q)} \text{ en } i\}$ ne sont pas pas disjoints, donc, en utilisant la relation $\mathbb{1}\{A\cup B\} = \mathbb{1}\{A\} + \mathbb{1}\{B\} - \mathbb{1}\{A\cap B\}$, on obtient

$$
\begin{aligned}
\widetilde{Y}_i^{(k)}(W) \;\; = \;\; & \sum_{C\in\mathcal{C}_k}\left(Y_i(C) - \mathbb{1}\left\{\bigcup_{p\in\mathcal{P}'(W)}\{W^{(p)}C \text{ en } i-p\}\right\} - \mathbb{1}\left\{\bigcup_{q\in\mathcal{P}'(W)}\{CW_{(q)} \text{ en } i\}\right\}\right.\\
& \left. +\mathbb{1}\left\{\bigcup_{p,q\in\mathcal{P}'(W)}\{W^{(p)}CW_{(q)} \text{ en } i-p\}\right\}\right).
\end{aligned}
$$

Chacune des réunions porte maintenant sur des événements disjoints, d'où le résultat énoncé par la proposition 4. □

La proposition précédente nous donne

$$
\widetilde{\mu}_k = \mathbb{E}(\widetilde{Y}_i^{(k)}) = \sum_{C\in\mathcal{C}_k}\mu(C) \;-2\sum_{C'\in\mathcal{C}_{k+1}}\mu(C') \;+\sum_{C''\in\mathcal{C}_{k+2}}\mu(C''). \tag{5.8}
$$

Remarques

1. Puisque $\widetilde{Y}_i = \sum_{k\geq 1}\widetilde{Y}_i^{(k)}$, on vérifie en sommant (5.8) sur l'ensemble d'indices $\{k\geq 1\}$ que $\widetilde{\mu} = \sum_{k\geq 1}\widetilde{\mu}_k$; en effet,

$$
\begin{aligned}
\sum_{k\geq 1}\widetilde{\mu}_k &= \sum_{k\geq 1,\, C\in\mathcal{C}_k}\mu(C) \;-2\sum_{k\geq 2,\, C\in\mathcal{C}_k}\mu(C) \;+\sum_{k\geq 3,\, C\in\mathcal{C}_k}\mu(C)\\
\sum_{k\geq 1}\widetilde{\mu}_k &= \sum_{C\in\mathcal{C}_1}\mu(C) \;-\sum_{C'\in\mathcal{C}_2}\mu(C')\\
&= \mu(W) \;-\sum_{p\in\mathcal{P}'(W)}\mu(W^{(p)}W)\\
&= \widetilde{\mu}. \qquad (5.9)
\end{aligned}
$$

La troisième égalité vient de $\mathcal{C}_1 = \{W\}$ et $\mathcal{C}_2 = \{W^{(p)}W,\ p\in\mathcal{P}'(W)\}$, et la dernière découle de la Proposition 3.

2. De la même façon, la formule (5.8) nous donne l'équation $\mu = \sum_{k\geq 1}k\,\widetilde{\mu}_k$ car

$$
\begin{aligned}
\sum_{k\geq 1}k\,\widetilde{\mu}_k &= \sum_{k\geq 1,\, C\in\mathcal{C}_k}\mu(C) \;-2\sum_{k\geq 2,\, C\in\mathcal{C}_k}(k-1)\mu(C) \;+\sum_{k\geq 3,\, C\in\mathcal{C}_k}(k-2)\mu(C)\\
&= \sum_{C\in\mathcal{C}_1}\mu(C)\\
&= \mu(W), \qquad (5.10)
\end{aligned}
$$

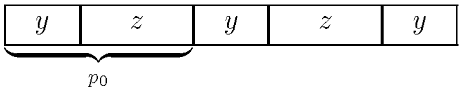
et en particulier $\mathbb{E}N(W) = \mathbb{E}N^*(W)$.

3. La formule (C.3) démontrée en Annexe met en évidence la décroissance géométrique du paramètre $\widetilde{\mu}_k$; en effet, on montre qu'il existe une variable $0 < A < 1$ explicite telle que

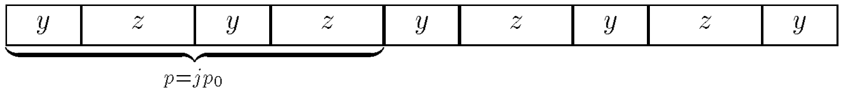
$$\widetilde{\mu}_k = (1 - A)^2 A^{k-1} \mu(W) .$$

Description des motifs composés

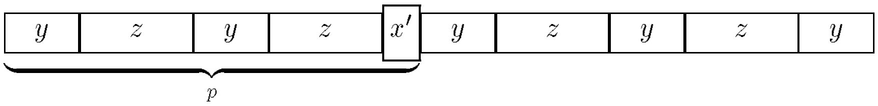
On est maintenant amené à décrire l'ensemble $\mathcal{C}_k$ des motifs composés, k étant fixé, pour permettre le calcul de $\mu(C)$ dans (5.8), lorsque $C \in \mathcal{C}_k$. Pour cela, nous donnons la structure d'un motif C composé de k occurrences chevauchantes du mot W. On considère la décomposition canonique $(yz)^r y$ de W ; en utilisant la description des périodes de W fournie par le théorème 10, on sait qu'il y a trois types de chevauchements illustrés ci-dessous :

W 

$(yz)^r y$

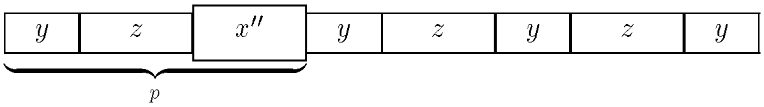
(i) 

$(yz)^{r+j} y$, $j \leq r$

(ii) 

$(yz)^r x'(yz)^r y$

où x' est une racine de y,

(iii) 

$(yz)^{r-1} x''(yz)^r y$

où x'' est une racine de yzy distincte de yz.

En remarquant que le chevauchement associé à la période jp_0 correspond à la composition de j chevauchements associés à la période minimale p_0 (chevauchements *triviaux*), le motif C s'écrit uniquement comme composition de chevauchements principaux. Par conséquent, une façon simple consiste à écrire C comme la concaténation de $(k-1)$ racines principales et d'un motif W :

$$C = W^{(p_1)} W^{(p_2)} \cdots W^{(p_{k-1})} W \,, \; p_i \in \mathcal{P}'(W) \,.$$

De façon plus précise, le schéma précédent montre que chacun des motifs yzx' (cas (ii)) et x'' (cas (iii)) est nécessairement précédé par $(yz)^{r-1}$, et que les puissances de yz strictement supérieures à r correspondent à des chevauchements triviaux. Ainsi, en utilisant le fait que les chevauchements principaux sont exclusifs, le motif C admet une unique décomposition : $C = (yz)^{r-1+\ell_1} x_1 (yz)^{r-1+\ell_2} x_2 \cdots x_{s-1} (yz)^{r+\ell_s} y$, que l'on écrira plus formellement

$$C = \prod_{i=1}^{s-1} \left((yz)^{r-1+\ell_i} x_i \right) \; (yz)^{r+\ell_s} y$$

où $s \in \mathbb{N}\backslash\{0\}$, $(\ell_1, \ldots, \ell_s) \in \mathbb{N}^s$ et les mots x_i s'écrivent soit $x_i = yzy'$, avec y' racine de y, soit $x_i = yz'$, avec yz' racine de yzy distincte de yz ; autrement dit x_i est une racine de yzy, différente de yz.

Or, un tel motif C contient exactement $\sum_{i=1}^{s}(l_i+1)$ mots W, ce qui nous permet d'établir la proposition suivante :

Proposition 5 *Soient un mot W, $(yz)^r y$ sa décomposition canonique et $\mathcal{C}_k$ l'ensemble de tous les motifs composés d'une succession de k occurrences chevauchantes de W. Alors, on a*

$$\mathcal{C}_k = \Big\{ \prod_{i=1}^{s-1} \Big((yz)^{r-1+\ell_i} x_i \Big) \; (yz)^{r+\ell_s} y \;\Big|\; s \geq 1,\ \forall\, i \in \{1, \ldots, s\},\ \ell_i \geq 0 \ \textit{et} \ \sum_{i=1}^{s}(l_i+1) = k,$$
$$\forall\, i \in \{1, \ldots, s-1\},\ x_i \ \textit{est une racine de } yzy \ \textit{distincte de } yz \Big\}.$$

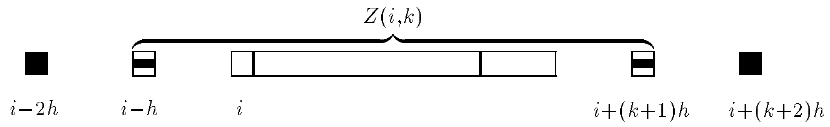
La longueur d'un motif $C \in \mathcal{C}_k$ est ainsi bornée : $h + (k-1)p_0 \leq |C| \leq kh - 2(k-1)$ ($< kh$ si $k > 1$).

Calcul des termes b_1^*, b_2^* et b_3^*

L'expression de ces trois quantités est donnée en 5.3.1 et leur calcul nécessite le choix d'un voisinage pour le double indice $(i,k) \in I^*$, où $I^* = I \times \mathbb{N}\backslash\{0\}$.

• On définit la zone de (i,k), notée $Z(i,k)$, comme un ensemble de positions j dans la séquence, qui contient au moins les indices j des lettres X_j qui définissent $\widetilde{Y}_i^{(k)}$. Comme on l'a vu, un k-motif est de longueur inférieure à kh, et il faut encore $(h-1)$ lettres avant i et après le k-motif pour s'assurer qu'il s'agit d'un k-train. On peut donc considérer les indices j tels que

$$Z(i,k) = \{j \text{ tel que } i - h \leq j \leq i + (k+1)h\}$$



Pour disposer, dans le terme b_3^*, d'un nombre de lettres de l'ordre de h entre une zone $Z(i,k)$ et une zone $Z(j,\ell)$, on définit (i,k) et (j,ℓ) comme non voisins si les deux zones associées sont distantes d'au moins h positions, c'est à dire si l'une des deux conditions suivantes sont réalisées :

$$\begin{aligned} i + (k+2)h &< j - h, \\ j + (\ell+2)h &< i - h. \end{aligned}$$

Le voisinage $B_{i,k}$ de (i,k) est défini comme l'ensemble des indices (j,ℓ) voisins de (i,k) :

$$B_{i,k} = \{(j,\ell) \text{ tels que } -(\ell+3)h \leq j - i \leq (k+3)h\} \cap I^*.$$

• Le terme b_1^* peut ainsi s'écrire

$$\begin{aligned} b_1^* &= \sum_{(i,k)\in I^*} \sum_{(j,\ell)\in B_{i,k}} \mathbb{E}(\widetilde{Y}_i^{(k)})\mathbb{E}(\widetilde{Y}_j^{(\ell)}) \\ &\leq \sum_{(i,k)\in I^*} \sum_{\ell\geq 1} \sum_{j=i}^{i+(k+3)h} \widetilde{\mu}_k\widetilde{\mu}_\ell + \sum_{(i,k)\in I^*} \sum_{\ell\geq 1} \sum_{j=i-(\ell+3)h}^{i-1} \widetilde{\mu}_k\widetilde{\mu}_\ell \end{aligned}$$

en séparant les indices j inférieurs à i de ceux supérieurs à i.
Puisque $\widetilde{\mu}_k\widetilde{\mu}_\ell$ ne dépend pas de j, par un argument de symétrie en k et ℓ des deux termes, on obtient la majoration

$$b_1^* \leq 2\sum_{i\in I}\left(\sum_{\ell\geq 1}\widetilde{\mu}_\ell\right)\left(\sum_{k\geq 1}((k+3)h+1)\widetilde{\mu}_k\right),$$

puis on utilise les relations (5.10) et (5.9), c'est à dire

$$\begin{aligned} \sum_{\ell\geq 1}\widetilde{\mu}_\ell &= \widetilde{\mu} \\ \sum_{k\geq 1}k\widetilde{\mu}_k &= \mu. \end{aligned}$$

Lorsqu'il n'y a pas d'ambiguïté nous écrivons μ pour $\mu(W)$ et $\widetilde{\mu}$ pour $\widetilde{\mu}(W)$.

Par conséquent,

$$b_1^* \leq 2(n-h+1)\,(h\mu + (3h+1)\widetilde{\mu})\,\widetilde{\mu}. \tag{5.11}$$

• Le terme b_2^* fait intervenir l'espérance des produits $\widetilde{Y}_i^{(k)}\widetilde{Y}_j^{(\ell)}$ lorsque $(j,\ell)\in B_{i,k}\backslash\{(i,k)\}$. Par symétrie du voisinage, on écrit b_2^* de la façon suivante

$$b_2^* \leq 2\sum_{(i,k)\in I^*}\sum_{\ell\geq 1}\sum_{j=i+1}^{i+(k+3)h}\mathbb{E}\left(\widetilde{Y}_i^{(k)}\widetilde{Y}_j^{(\ell)}\right).$$

On remarquera que le terme correspondant à $j = i$ a disparu car deux trains de longueur différente ne peuvent pas commencer simultanément au même endroit ; ainsi $\widetilde{Y}_i^{(k)}\widetilde{Y}_i^{(\ell)} = 0$ si $k \neq \ell$.

Une majoration directe de b_2^* n'assure pas la convergence de b_2^* vers 0. En effet, en sommant l'espérance $\mathbb{E}\left(\widetilde{Y}_i^{(k)}\widetilde{Y}_j^{(\ell)}\right)$ sur $\{\ell\geq 1\}$, on obtient $\mathbb{E}\left(\widetilde{Y}_i^{(k)}\widetilde{Y}_j\right)$ que l'on peut majorer sans risque par $\mathbb{E}\left(\widetilde{Y}_i^{(k)}Y_j\right)$. Ensuite, soit on majore brutalement cette espérance par $\mathbb{E}\left(\widetilde{Y}_i^{(k)}\right)$ de façon à faire apparaître un terme $h\mu$ lorsque l'on effectue les sommes sur k et j, mais on perd un terme de l'ordre de μ dans le majorant de b_2^* qui devient un $\mathrm{O}(nh\mu)$, soit on majore $\mathbb{E}\left(\widetilde{Y}_i^{(k)}Y_j\right)$, comme Geske *et al.* (1995), par $\sum_{C\in\mathcal{C}_k}\mu(C)\times\mu(W)$, ce qui nous donne la majoration $b_2^*\leq 2nh\mu\sum_k(k+3)\sum_{C\in\mathcal{C}_k}\mu(C)$ et en utilisant le fait que $\mu(C)\leq\mu A^{k-1}$ où A est une certaine

quantité plus petite que 1, on obtient un majorant pour b_2^* de l'ordre de $nh\mu^2 \sum_k k.\text{card}(\mathcal{C}_k).A^{k-1}$ qui peut diverger car $\text{card}(\mathcal{C}_k)$ $=\text{card}(\mathcal{P}'(W))^{k-1}$. Le seul cas où l'on peut conclure de façon positive est celui pour lequel $\text{card}(\mathcal{P}'(W)) = 1$, ce qui est exactement le cas dans lequel se placent Geske *et al.* (1995).
La majoration par $\sum_{C\in\mathcal{C}_k} \mu(C)$ est trop forte ici car nous perdons le fait qu'un seul motif C peut apparaître en i. De plus, certains produits $\widetilde{Y}_i^{(k)}\widetilde{Y}_j^{(\ell)}$ sont nuls puisque le train qui commence en j ne peut chevaucher, par définition, celui qui commence en i. Nous devons les identifier afin de prouver la convergence de b_2^* vers zéro.

Pour identifier ces produits nuls, il nous faut connaître les motifs composés C et C' qui apparaissent en i et j ; par conséquent, comme nous avions décomposé $\widetilde{Y}_i$ en (5.5), nous décomposons $\widetilde{Y}_i^{(k)}$ de la façon suivante :

$$\widetilde{Y}_i^{(k)} = \sum_{\substack{U\in\mathcal{U}\\ C\in\mathcal{C}_k\\ V\in\mathcal{V}}} Y_{i-h}(UCV), \tag{5.12}$$

où $\mathcal{U}$ est défini en (5.4) et $\mathcal{V}$ est l'ensemble des mots de longueur h qui ne commencent pas par un suffixe $W_{(p)}$ avec p période principale, autrement dit

$$\mathcal{V} = \{v_1\cdots v_h \mid \forall p\in\mathcal{P}'(W),\ v_1\cdots v_p \neq W_{(p)}\}. \tag{5.13}$$

En utilisant la même décomposition pour $\widetilde{Y}_j^{(\ell)}$ avec $U'\in\mathcal{U}$, $C'\in\mathcal{C}_\ell$ et $V'\in\mathcal{V}$, on a

$$b_2^* = 2\sum_{(i,k)\in I^*}\ \sum_{U,C,V}\ \sum_{\ell\geq 1}\ \sum_{U',C',V'}\ \sum_{j=i+|C|}^{i+(k+3)h} \mathbb{E}\left(Y_{i-h}(UCV)Y_{j-h}(U'C'V')\right)\ ;$$

dans cette formule, les indices j inférieurs à $i+|C|$ n'interviennent plus car $i+1\leq j< i+|C|$ signifie que le motif composé C' en j chevauche C en i, ce qui est impossible. On distingue deux termes b_{21}^* et b_{22}^* qui correspondent au cas où les motifs UCV et $U'C'V'$ se chevauchent ou non.
— Lorsqu'ils ne se chevauchent pas, le terme b_{22}^* s'écrit

$$b_{22}^* \leq 2\sum_{(i,k)\in I^*}\ \sum_{U,C,V}\ \sum_{\ell\geq 1}\ \sum_{U',C',V'}\ \sum_{j=i+|C|+2h}^{i+(k+3)h} \frac{\mu(UCV)\mu(U'C'V')}{\mu(u_1')}.$$

car $\mathbb{E}\left(Y_{i-h}(UCV)Y_{j-h}(U'C'V')\right) \leq \mu(u_1')^{-1}\mu(UCV)\mu(U'C'V')$, où u_1' est la première lettre de U'.
La somme sur j contient $(k+3)h-|C|-2h+1$ termes qui ne dépendent pas de j ; de plus, en minorant $|C|$ par h et $\mu(u_1')$ par μ_{inf}, puisque $\mu(u_1')$ ne peut prendre que quatre valeurs dont la plus petite est notée μ_{inf}, on obtient

$$b_{22}^* \leq \frac{2}{\mu_{\text{inf}}}\sum_{i\in I}\left(\sum_{k\geq 1}(kh+1)\sum_{U,C,V}\mu(UCV)\right)\left(\sum_{\ell\geq 1}\ \sum_{U',C',V'}\mu(U'C'V')\right)\ ;$$

or

$$\begin{aligned}\sum_{\ell\geq 1}\sum_{U',C',V'}\mu(U'C'V') &= \sum_{\ell\geq 1}\widetilde{\mu}_\ell=\widetilde{\mu},\\ \sum_{k\geq 1}k\sum_{U,C,V}\mu(UCV) &= \sum_{k\geq 1}k\widetilde{\mu}_k=\mu,\end{aligned}$$

d'où

$$b^*_{22}\leq\frac{2}{\mu_{\inf}}(n-h+1)(h\mu+\widetilde{\mu})\widetilde{\mu}. \tag{5.14}$$

— Le terme b^*_{21} qui correspond à deux motifs UCV et $U'C'V'$ qui se chevauchent sur au plus $2h$ lettres, s'écrit

$$b^*_{21}=2\sum_{(i,k)\in I^*}\sum_{U,C,V}\sum_{\ell\geq 1}\sum_{U',C',V'}\sum_{j=i+|C|}^{i+|C|+2h-1}\mathbb{E}\left(Y_{i-h}(UCV)Y_{j-h}(U'C'V')\right),$$

c'est à dire, en effectuant la somme sur ℓ et celle sur U',C',V',

$$b^*_{21}=2\sum_{(i,k)\in I^*}\sum_{U,C,V}\sum_{j=i+|C|}^{i+|C|+2h-1}\mathbb{E}\left(Y_{i-h}(UCV)\widetilde{Y}_j(W)\right).$$

On majore ensuite $\widetilde{Y}_j(W)$ par $Y_j(W)$ et $\sum_{V\in\mathcal{V}}Y_{i-h}(UCV)$ par $Y_{i-h}(UC)$, ce qui nous donne

$$b^*_{21}\leq 2\sum_{(i,k)\in I^*}\sum_{U,C}\sum_{j=i+|C|}^{i+|C|+2h-1}\mathbb{E}\left(Y_{i-h}(UC)\,Y_j(W)\right).$$

On utilise, alors, le fait que les motifs UC en $i-h$ et W en j ne se chevauchent pas, pour obtenir l'inégalité suivante

$$\mathbb{E}\left(Y_{i-h}(UC)\,Y_j(W)\right)\leq\frac{\mu(UC)\mu(W)}{\mu(u)}\ ;$$

de cette façon,

$$b^*_{21}\leq\frac{4}{\mu(u)}\,h\,\mu(W)\sum_{(i,k)\in I^*}\sum_{U,C}\mu(UC).$$

Pour finir, il suffit de remarquer que $\sum_k\sum_{U,C}\mu(UC)=\mu(W)$; en effet,

$$\begin{aligned}\sum_{U\in\mathcal{U},C\in\mathcal{C}_k}\mu(UC) &= \mathrm{P}\{\text{un train d'au moins } k \text{ occurrences commence en } i\}\\ &= \sum_{K\geq k}\widetilde{\mu}_K,\end{aligned}$$

d'où

$$\sum_{k\geq 1}\sum_{U\in\mathcal{U},C\in\mathcal{C}_k}\mu(UC)=\sum_{k\geq 1}\sum_{K\geq k}\widetilde{\mu}_K=\sum_{K\geq 1}K\,\widetilde{\mu}_K=\mu(W).$$

Ainsi, on obtient la majoration

$$b_{21}^* \leq \frac{4}{\mu(u)}(n-h+1)h\mu^2(W). \tag{5.15}$$

• Enfin, la majoration du terme b_3^* se calque exactement sur celle de $b_3(\widetilde{Y})$ (formule (5.6)), à cela près que l'on utilise la décomposition (5.12), et on obtient

$$b_3^* \leq C\,(n-h+1)\,\widetilde{\mu}(W)|\alpha|^h \tag{5.16}$$

où C est une constante et α est la seconde plus grande valeur propre, en module, de la matrice de transition (par conséquent $|\alpha| < 1$). L'exposant h de cette valeur propre provient du choix du voisinage $B_{i,k}$ que l'on a fait auparavant.

En reprenant l'équation d'approximation (5.7) et les majorations (5.11), (5.14), (5.15) et (5.16), on établit le théorème suivant :

Théorème 14 *Soit W un mot de longueur h. On a*

$$d_{VT}(\mathcal{L}(N(W)), CP(\Lambda)) \leq (n-h+1)\left(A_1 h\,\mu^2 + A_2 h\,\mu\widetilde{\mu} + A_3 h\,\widetilde{\mu}^2 + A_4\widetilde{\mu}^2 + A_5\widetilde{\mu}|\alpha|^h\right) + 4h\,\mu$$

où les quantités A_1, A_2, A_3, A_4 et A_5 sont des O(1) et $CP(\Lambda)$ représente la loi de Poisson composée de $\sum_{k\geq 1} k\,Z^{(k)}$ où les $Z^{(k)}$ sont des variables aléatoires indépendantes distribuées suivant la loi de Poisson $\mathcal{P}\left((n-h+1)\widetilde{\mu}_k\right)$. L'expression de $\widetilde{\mu}_k$ est donnée en (5.8).

Corollaire 2 *On considère une suite de mots W_n de longueur h_n telle que $n\mu(W_n) = O(1)$ et $h_n = o(n)$. On a*

$$\lim_{n\to+\infty} d_{VT}(\mathcal{L}(N(W_n)), CP(\Lambda_n)) = 0\,,$$

où $CP(\Lambda_n)$ est la loi de Poisson composée, définie au théorème 14, pour le mot W_n.

Comparaison avec les résultats de Geske *et al.* (1995)

– Geske *et al.* (1995) ont démontré un résultat analogue, en utilisant de la même façon la méthode de Chen-Stein, mais uniquement pour des mots W ayant une seule période principale, c'est à dire la période minimale. En effet, après avoir défini la notion de trains comportant exactement k occurrences de W, ils en donnent la condition équivalente suivante : *le motif composé, associé au train, doit être encadré à droite et à gauche par deux mots de longueur p_0 (la période minimale) qui ne permettent pas d'étendre ce motif composé.* Or, comme nous l'avons vu, pour un mot quelconque cette condition n'est pas suffisante car il faut tenir compte des autres périodes principales (proposition 4). Si nous prenons l'exemple du mot $W =$ `AACAA`, il ne suffit pas d'exclure, devant (par exemple) une occurrence de W, la racine minimale `AAC`, il faut aussi exclure la racine AACA. Par conséquent, ils se restreignent à une classe simple de mots : ceux ne possédant qu'une période principale. De cette façon, les motifs C, composés de k occurrences d'un mot W, qu'ils étudient sont d'autant plus simples que C est unique et de la forme $C = (yz)^{r+k-1}y$ si $(yz)^r y$ est la décomposition canonique de W. Autrement dit, pour tout $k \geq 1$, leur ensemble $\mathcal{C}_k$ est de cardinal égal à 1, alors qu'ici ces cardinaux sont de l'ordre de β^k où β est le nombre de périodes principales de W, ce qui paraît beaucoup plus pénalisant puisque k décrit $\mathbb{N}\backslash\{0\}$. De plus, dans le cas général, la structure des motifs composés (proposition 5) est bien plus complexe que la structure périodique $(yz)^{r+k-1}y$. À ce sujet, il semble inaccessible de pouvoir déterminer le nombre maximal d'occurrences non chevauchantes de W dans de tels motifs, ce qu'ils ont pu faire aisément dans leur cas. Dans leurs majorations des quantités b_1^*, b_2^* et b_3^*, ils ont utilisé le fait que $\widetilde{\mu}_k \leq \mu\, A^{k-1}$ où $A = \pi(w_1, w_2)\cdots\pi(w_{p_0}, w_{p_0+1})$, d'où les inégalités suivantes :

$$\sum_{k\geq 1} \widetilde{\mu}_k \leq \frac{\mu}{1-A} \qquad \text{et} \qquad \sum_{k\geq 1} k\widetilde{\mu}_k \leq \frac{\mu}{(1-A)^2},$$

et ont supposé que A est majoré uniformément par $1-\delta$ $(0 < \delta < 1)$ ce qui, il est vrai, est raisonnable. Or nous avons montré dans le cas général, c'est à dire pour un mot W quelconque, que

$$\sum_{k\geq 1} \widetilde{\mu}_k \leq \mu \qquad \text{et} \qquad \sum_{k\geq 1} k\widetilde{\mu}_k = \mu.$$

Le coût de cette généralisation a été essentiellement un travail assez conséquent de combinatoire sur les mots.

– Nous obtenons une erreur d'approximation du même ordre que celle obtenue par Geske *et al.* (1995) dans un cas particulier. Comme ils le font remarquer, la structure des voisinages utilisés pour appliquer la méthode de Chen-Stein n'est pas rectangulaire, contrairement à ce qui est fait dans Arratia *et al.* (1990b), par exemple. Cependant, nous ne faisons pas intervenir l'ordre de la seconde valeur propre de la matrice de transition dans la définition de nos voisinages, comme le font Geske *et al.* (1995) ; cette valeur propre apparaît tout de même dans l'expression du majorant de l'erreur d'approximation.

5.3.4 Généralisation au modèle markovien d'ordre m : Mm

Pour généraliser le travail précédent au cas où l'ordre du modèle est quelconque, il suffit de remarquer que si la séquence $S = X_1X_2X_3\cdots X_n$ est une chaîne de Markov d'ordre m sur l'alphabet $\mathcal{A}$, alors la séquence $\mathbb{S} = \mathbb{X}_1\mathbb{X}_2\mathbb{X}_3\cdots\mathbb{X}_{n-m+1}$ définie par $\mathbb{X}_i = X_iX_{i+1}\cdots X_{i+m-1}$ est une chaîne de Markov d'ordre 1 sur l'alphabet $\mathcal{A}^m$. Aussi, un mot $W = w_1w_2\cdots w_h$ écrit dans l'alphabet $\mathcal{A}$, s'écrit $\mathbb{W} = W_1W_2\cdots W_{h-m+1}$ dans $\mathcal{A}^m$ avec $W_j = w_jw_{j+1}\cdots w_{j+m-1}$. Par construction, si un mot W commence en position i dans la séquence S alors un mot $\mathbb{W}$ commence en position i dans la séquence $\mathbb{S}$, d'où

$$\mathbb{Y}_i \stackrel{\text{déf}}{=} \mathbb{1}\{\text{un mot } \mathbb{W} \text{ commence en position } i \text{ dans } \mathbb{S}\} = Y_i\,.$$

Par conséquent, le nombre d'occurrences de W dans S est identique au nombre d'occurrences de $\mathbb{W}$ dans $\mathbb{S}$

$$N(W)[S] = N(\mathbb{W})[\mathbb{S}]\,,$$

que l'on notera sans ambiguïté $N(W)$ et $N(\mathbb{W})$. Pourvu que la séquence $\mathbb{S}$ soit assez longue et le mot $\mathbb{W}$ suffisamment grand, par le théorème 14 on approche la loi de $N_{\mathbb{W}}$ par la loi de Poisson composée CP($\mathbb{L}$) où la mesure $\mathbb{L}$ est définie par $\mathbb{L} \equiv \sum_k \delta_k \widetilde{\mathbb{L}}_k$ avec

$$\widetilde{\mathbb{L}}_k = (n-h+1)\,\mathbb{E}\widetilde{\mathbb{Y}}_i^{(k)}$$

et

$$\widetilde{\mathbb{Y}}_i^{(k)} = \mathbb{1}\{\text{un train composé de } k \text{ occurrences de } \mathbb{W} \text{ commence en } i \text{ dans } \mathbb{S}\}.$$

Si on note $\mathbb{C}_k$ l'ensemble des motifs composés de k occurrences de $\mathbb{W}$, on obtient par la proposition (4) :

$$\mathbb{E}\widetilde{\mathbb{Y}}_i^{(k)} = \sum_{\mathbb{C}\in\mathbb{C}_k} \mu(\mathbb{C}) - 2 \sum_{\mathbb{C}'\in\mathbb{C}_{k+1}} \mu(\mathbb{C}') + \sum_{\mathbb{C}''\in\mathbb{C}_{k+2}} \mu(\mathbb{C}'')\,,$$

où $\mu(\mathbb{V})$ est définit par $\mu(\mathbb{V}) = \mathbb{E}\mathbb{Y}_i(\mathbb{V}) = \mu(V_1)\prod_{i=1}^{|\mathbb{V}|-1}\pi(V_i, V_{i+1})$ si $\mathbb{V} = V_1V_2\cdots V_{|\mathbb{V}|}$.

Pour obtenir une approximation de la loi de $N(W)$, nous reformulons, dans l'alphabet $\mathcal{A}$, l'expression de $\mathbb{E}\widetilde{\mathbb{Y}}_i^{(k)}$ ci-dessus. En effet, les périodes de $\mathbb{W}$, dont l'ensemble est noté $\mathcal{P}(\mathbb{W})$, sont celles de W qui sont strictement inférieures à $h-m+1$, la longueur du mot $\mathbb{W}$:

$$\mathcal{P}(\mathbb{W}) = \mathcal{P}(W) \bigcap \{1,\ldots,h-m\}\,,$$

et il en est de même pour les périodes principales :

$$\mathcal{P}'(\mathbb{W}) = \mathcal{P}'(W) \bigcap \{1,\ldots,h-m\} \stackrel{\text{déf}}{=} \mathcal{P}'_{(m)}(W).$$

Donc, si $p \in \mathcal{P}'(\mathbb{W})$, on note $\mathbb{W}^{(p)}$ la racine principale associée et le mot composé $\mathbb{W}^{(p)}\mathbb{W}$ s'écrit, dans l'alphabet $\mathcal{A}$, $W^{(p)}W$ avec $p \in \mathcal{P}'_{(m)}(W)$.
Ainsi, l'ensemble $\mathbb{C}_k$ correspond, dans l'alphabet $\mathcal{A}$, à l'ensemble $\mathcal{C}_k^{(m)}$ défini comme l'ensemble des motifs composés appartenant à $\mathcal{C}_k$ et construits uniquement à partir des périodes appartenant à $\mathcal{P}'_{(m)}(W)$.

Par conséquent, l'expression de $\widetilde{\mathbb{L}}_k = (n-h+1)\mathbb{E}\widetilde{\mathbb{Y}}_i^{(k)}$ sur l'alphabet $\mathcal{A}$ est

$$\widetilde{\mathbb{L}}_k = (n-h+1)\left(\sum_{C\in\mathcal{C}_k^{(m)}} \mu_m(C) - 2\sum_{C'\in\mathcal{C}_{k+1}^{(m)}} \mu_m(C') + \sum_{C''\in\mathcal{C}_{k+2}^{(m)}} \mu_m(C'')\right)\,, \tag{5.17}$$

où $\mu_m(V)$ est définit par $\mu_m(V) = \mathbb{E}Y_i(V) = \mu(v_1\cdots v_m)\prod_{i=1}^{|V|-m}\pi(v_i\cdots v_{i+m-1}, v_{i+m})$ si $V = v_1v_2\cdots v_{|V|}$.

Par suite on obtient le théorème suivant :

Théorème 15 *On considère une suite de mots W_n de longueur h_n telle que $n\mu_m(W_n) = O(1)$ et $h_n - m + 1 = o(n)$. On a*

$$\lim_{n\to+\infty} d_{VT}\left(\mathcal{L}(N(W_n)), CP\left(\sum_k \delta_k\,\widetilde{\mathbb{L}}_k\right)\right) = 0$$

où les paramètres $\widetilde{\mathbb{L}}_k$ se déduisent de (5.17) pour le mot W_n.

Remarques

1. On n'obtient pas ici de résultats sur la loi de $\widetilde{N}^*(W)$ car en général $\widetilde{N}^*(W) \neq \widetilde{N}^*(\mathbb{W})$; en effet, les trains de $\mathbb{W}$ ne font pas intervenir les chevauchements de W qui correspondent à des périodes $h - m < p < h$.

2. La loi de Poisson composée limite se réduit à une loi de Poisson si $\mathcal{P}'_{(m)}(W) = \emptyset$, autrement dit si le mot W n'a pas de périodes principales inférieures à $h - m$.

Troisième partie

Application : recherche de mots de fréquence exceptionnelle dans des séquences d'ADN

Chapitre 6

Présentation

6.1 Intérêt biologique

La principale motivation du travail présenté dans les deux parties précédentes était de proposer à nos collègues biologistes une méthode statistique qui réponde à une de leurs questions : "comment détecter automatiquement des oligonucléotides (mots) qui apparaissent avec une fréquence significativement exceptionnelle dans une séquence d'ADN ?". Cette question est importante car les mots de fréquence exceptionnelle peuvent correspondre à des fonctions biologiques qu'il serait intéressant de découvrir. On connaît par exemple, des mots dont la fréquence élevée ou faible est liée à la stabilité du génome.

Ainsi, une étude approfondie de la fréquence du motif GCTGGTGG dans une longue séquence d'ADN de la bactérie *E. coli* est présentée au Chapitre 9 car ce motif, appelé *chi* de *E. coli*, est largement représenté le long du génome de la bactérie afin de le protéger de la dégradation d'une enzyme spécifique.

De la même façon, une étude systématique sur le génome de *E. coli* des mots palindromiques, qui se lisent de la même façon en sens inverse sur le brin complémentaire, est présentée au Chapitre 8, car les enzymes de restriction des bactéries sont capables de couper les brins d'ADN en des sites de restriction qui sont bien souvent des palindromes. Il semble ainsi naturel que de tels sites soient peu représentés sur le génome bactérien. On présentera également les conclusions d'une étude parallèle faite sur la bactérie *B. subtilis*.

L'étude de la fréquence du chi de *E. coli*, ainsi que celle des palindromes seront aussi réalisées sur les génomes de deux bactériophages de *E. coli*, parasites de la bactérie : le phage *Lambda* et le phage *T7*. Si ces deux mécanismes cités ci-dessus sont un moyen pour la bactérie de se défendre contre toute intrusion de génome étranger, on peut se demander si les bactériophages se seraient "adaptés" pour survivre dans la bactérie. Pour ces parasites, un moyen de survivre dans *E. coli* pourrait être en effet d'avoir évolué de façon à ce que leur génome possède un nombre important de motifs GCTGGTGG et au contraire très peu de sites de restriction.

Enfin, le Chapitre 7 est consacré à l'étude des petits mots phasés de longueur 3 et 4, et en particulier des codons, d'une séquence codante de *E. coli* puis de *B. subtilis* avec les modèles périodiques d'ordre 1 et 2. Il ne paraît pas surprenant, du fait de la contrainte de la traduction en protéine des séquences codantes, de constater que les codons possèdent des fréquences exceptionnelles. Pour certains acides aminés, une préférence de codon est aussi constatée. Il apparaît clairement que le modèle *M2_3* (ou même *M1_c*) qui s'ajuste sur la fréquence des codons est le modèle simple et adéquat pour l'étude des séquences codantes.

6.2 Caractère exceptionnel d'un mot

Un mot exceptionnel, ou plutôt de fréquence exceptionnelle, dans une séquence est un mot sous- ou sur-représenté par rapport à ce que l'on attend sous un modèle. Le choix et l'influence du modèle sont bien entendu important pour l'analyse d'une séquence, et seront développés au paragraphe suivant.

En se plaçant dans un modèle M, on peut définir un mot sur-représenté (resp. sous-représenté) comme un mot W dont le comptage $N^{\text{obs}}(W)$ observé dans la séquence se situe dans la queue droite (resp. gauche) de la distribution de $N(W)$ calculée sous le modèle M. On calculera donc les P-valeurs $\mathrm{P}\{N(W) > N^{\text{obs}}(W)\}$ pour chaque mot W. On a vu dans les parties précédentes que, suivant le cadre asymptotique choisi, la distribution de $N(W)$ pouvait être approchée soit par une loi normale, soit par une loi de Poisson si W n'est pas chevauchant, soit par une loi de Poisson composée si W est chevauchant.

Afin de ne pas manipuler des queues de distribution proches de 0 ou de 1 pour ce qui est des mots qui nous intéressent, à savoir les mots exceptionnels, nous représenterons les résultats en échelle gaussienne standard, c'est à dire que :

- pour l'approximation gaussienne, on calcule pour chaque mot la statistique observée associée, qui est asymptotiquement gaussienne centrée réduite sous le modèle M (Partie I) :
$$U_M(W) = \frac{N^{\text{obs}}(W) - n\widehat{\mu}_M}{\widehat{\sigma}_M}$$
où $n\widehat{\mu}_M$ est l'estimateur naturel de $N(W)$ sous le modèle M, et $\widehat{\sigma}^2_M$ est un estimateur consistant de la variance conditionnelle asymptotique de $N(W)$ sous le modèle M, calculés en Partie I.
Remarque : On a choisi dans cette partie d'illustrer uniquement l'approche conditionnelle qui permet d'identifier des mots de fréquence exceptionnelle sans privilégier de lettres particulière dans le mot. Si l'on prend l'exemple d'un mot *abcd*, elle compare en quelque sorte N_{abcd} à $\frac{N_{ab}N_{bc}N_{cd}}{N_b N_c}$ et fait intervenir ainsi la composition globale du mot. L'approche martingale privilégie au contraire la dernière (ou la première) lettre du mot en comparant N_{abcd} à $\frac{N_{abc}N_{cd}}{N_c}$. On détecte ainsi si la présence de la lettre d après *abc* est exceptionnelle.

- Pour les approximations de Poisson, on calcule pour chaque mot le quantile de la gaussienne centrée réduite associé à la queue de distribution de la loi de Poisson ou Poisson composée ad-hoc. Ces queues sont en fait calculées pour les lois limites dont les paramètres sont estimés de façon naturelle dans le modèle M. En effet, en considérant par exemple la loi de Poisson, la distance en variation totale entre la loi de Poisson de paramètre $n\mu(W)$ et la loi de Poisson de paramètre $n\widehat{\mu}_M(W)$ est un $\mathrm{O}(n|\mu(W) - \widehat{\mu}_M(W)|)$ (Annexe C). On montre que cette quantité tend vers zéro pour l'estimateur naturel $n\widehat{\mu}_M(W) = N(w_1)\prod_{i=1}^{h-1} N(w_iw_{i+1})/N(w_i)$, en utilisant la loi du log-itéré et l'hypothèse asymptotique $n\mu(W) = \mathrm{O}(1)$

Un calcul plus fin de la queue de distribution peut être directement obtenu pour les mots exceptionnels à l'aide des techniques de grandes déviations pour les chaînes de Markov. Cela peut être un moyen de comparer l'approximation de la queue de distribution de $N(W)$ obtenue avec les lois limites gaussienne et de Poisson composée (Section 6.4).

6.3 Choix et influence du modèle

Se placer dans un modèle pour analyser une séquence est une façon de tenir compte de certaines caractéristiques de la séquence. Il est clair que si une séquence est riche en A, cela a tendance à favoriser une richesse en mots composés de A. Par conséquent, les mots exceptionnels ne sont pas nécessairement les mots qui possèdent les fréquences les plus fortes ou les plus faibles. En voici une illustration convaincante : dans une séquence de *E. coli* de longueur 500042 bp, en utilisant l'approximation gaussienne, GCGCC est le 5-mot le plus sous-représenté dans le modèle *M3* et est présent un nombre de fois plus élevé (954 occurrences) que TCCGG (738 occurrences) qui est le 5-mot le plus sur-représenté dans le même modèle.

Le choix du modèle, et plus particulièrement le choix de l'ordre m du modèle markovien, est ici le choix de l'information sur la séquence, conditionnellement à laquelle on souhaite étudier statistiquement la fréquence de certains mots de longueur au moins égales à $m+2$. Cette information sur la séquence est en fait sa composition en mots de longueur $1, 2, \ldots$ et $m+1$, car la statistique exhaustive du modèle markovien d'ordre m est constituée des m premières bases de la séquence et de l'ensemble des comptages des $(m+1)$-mots, phasés si l'on considère un modèle périodique. Les mots exceptionnels ne sont pas les mêmes d'un modèle à l'autre puisque lorsque l'on augmente l'ordre du modèle, on s'ajuste entre autres sur les fréquences exceptionnelles de mots détectées dans les modèles d'ordre inférieurs. Ainsi, le modèle maximal d'ordre $h-2$ est idéal pour identifier des mots de longueur h exceptionnels, puisque l'on s'ajuste sur les mots de longueur $h-1$. Seulement, outre le nombre de paramètres important du modèle maximal, on risque de ne pas identifier un h-mot qui, dans un modèle d'ordre inférieur, est exceptionnel par "contamination" d'un de ses sous-mots exceptionnel dans ce modèle.

L'étude des 4-mots réalisée ci-dessous sur une séquence de *E. coli* (500042 bp), illustre différents cas de figure quant à l'influence du modèle sur l'interprétation de la fréquence exceptionnelle d'un mot.
Chaque point de la Figure 6.1 représente un mot de longueur 4 et possède, comme abscisse, la statistique asymptotiquement gaussienne U_1 calculée dans le modèle *M1* pour le mot associé, et comme ordonnée, la statistique U_2 calculée dans le modèle *M2* . Cette figure nous permet non seulement de comparer le comportement exceptionnel des 4-mots dans les modèles *M1* et *M2* , mais aussi de comparer les mots entre eux.

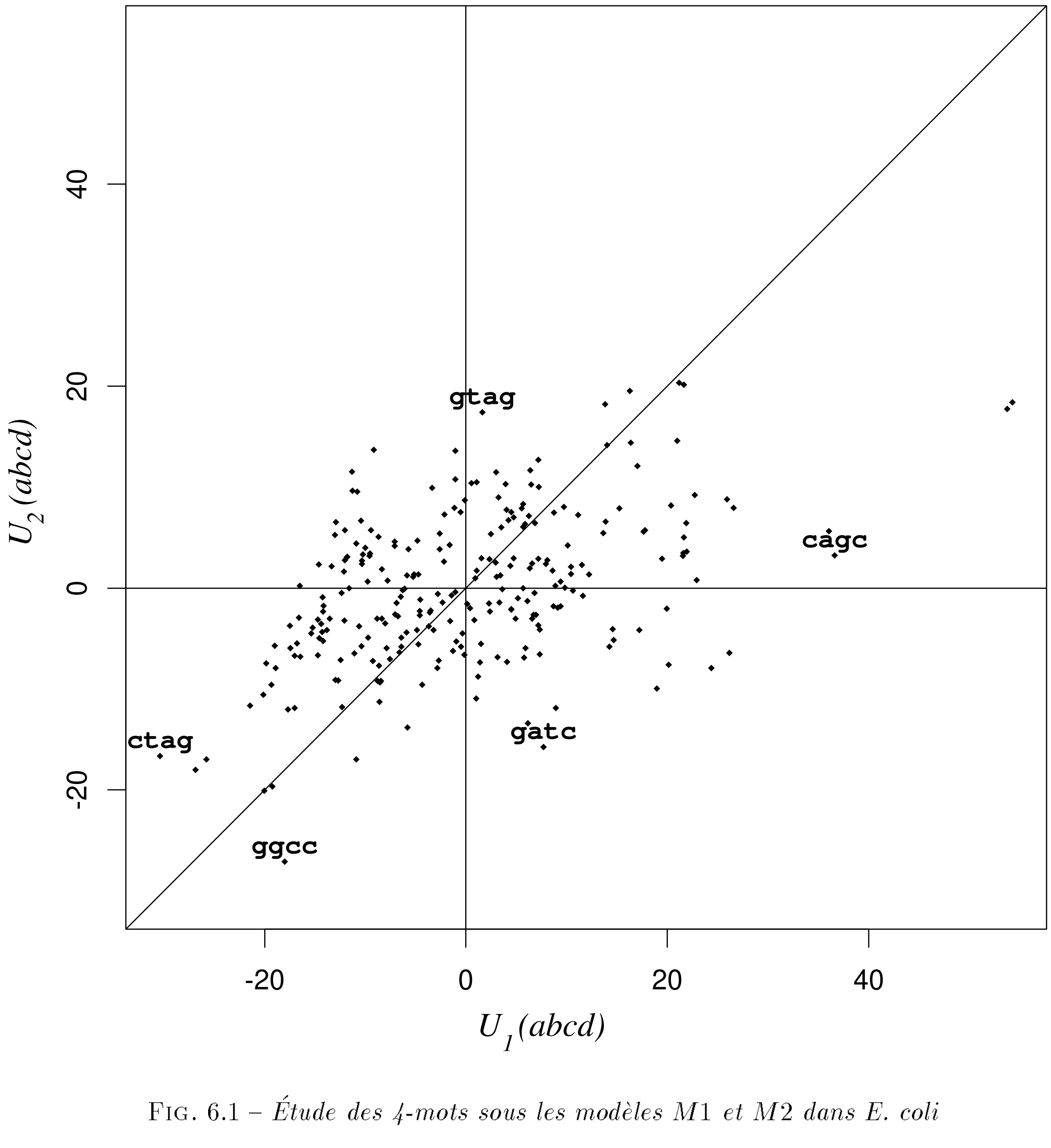


Fig. 6.1 – *Étude des 4-mots sous les modèles M1 et M2 dans E. coli*

Nous allons étudier successivement les 5 mots dont le nom est inscrit sur la Figure 6.1, mais auparavant nous représentons, pour chacun d'eux, leur pyramide dans le modèle *M1* (Figure 6.2.a), ainsi que celles dans les modèles maximaux. (Figure 6.2.b). Cette représentation pyramidale permet de visualiser simultanément le comportement statistique de la fréquence d'un mot ainsi que celui de ses sous-mots, soit dans un même modèle, soit dans le modèle d'ordre maximal associé à chaque longueur de mots.
En effet, la pyramide d'un mot de longueur 4 comporte deux étages : l'étage supérieur est constitué d'un carré dont l'intensité de gris est d'autant plus grande que la valeur de la statistique du mot de 4 lettres est grande (en valeur absolue) ; l'étage du dessous est constitué de deux carrés correspondant aux deux sous-mots de longueur 3. La pyramide d'un mot est surmontée d'un signe + ou - pour indiquer si la statistique du mot est positive ou négative. Par ailleurs, les sous-mots dont la statistique est de signe opposé à celle du mot au sommet, sont marqués du signe O placé au centre de leur carré.
Un phénomène de contamination est donc clairement détecté par cette représentation.

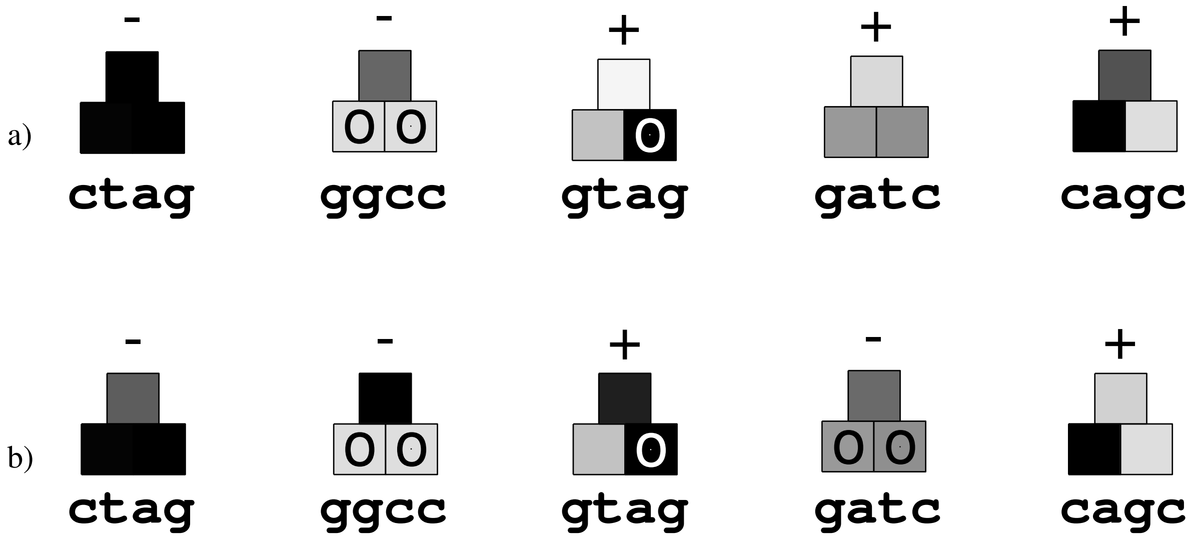


FIG. 6.2 – *Pyramides dans le modèle M1 (a), puis dans les modèles d'ordre maximal M2 et M1 (b).*

- TAG et CTA sont les deux tri-nucléotides les plus exceptionnellement rares dans le modèle *M1*. Il n'est donc pas surprenant que CTAG soit le tétra-nucléotide le plus exceptionnellement rare dans ce même modèle (Tableau 6.1). Remarquons tout de même, que parmi les 2822 CTA, seuls 94 sont suivis d'un G (même remarque pour les TAG précédés d'un C). Ainsi, le fait que CTA et TAG soient déjà peu représentés, ne suffit pas pour expliquer à quel point CTAG est évité.

  Pour pouvoir juger la significativité d'un tel phénomène, il faut se mettre dans le modèle

Tab. 6.1 – Statistiques de CTA, TAG et CTAG dans M1

W	$N(W)$	$n\widehat{\mu}_1(W)$	$\widehat{\sigma}_1^2(W)$	$U_1(W)$	Rang
TAG	2879	4727.326	3032.939	-33.5619	1
CTA	2822	4627.576	2988.420	-33.0290	2
CTAG	94	1004.456	897.058	-30.3982	1

Tab. 6.2 – Statistiques des huit tétra-nucléotides les plus rares dans M2

W	$N(W)$	$n\widehat{\mu}_2(W)$	$\widehat{\sigma}_2^2(W)$	$U_2(W)$	Rang
GGCC	1478	2503.995	1433.435	-27.0992	1
CCAA	1386	2071.047	1164.197	-20.0774	2
TTGG	1553	2249.444	1255.129	-19.6581	3
CTTG	1017	1579.316	976.558	-17.9942	4
CAAG	1035	1566.496	979.147	-16.9854	5
CGCG	3244	3986.334	1913.385	-16.9706	6
CTAG	94	373.113	281.820	-16.6262	7
GATC	2163	2744.382	1367.946	-15.7191	8

M2, qui s'ajuste sur le comptages des tri-nucléotides ; de cette façon, les déficiences en CTA et TAG seront intégrées dans le modèle et ne pourront donc plus apparaître comme responsables d'un écart au modèle *M2*. Si CTAG n'est plus le tétra-nucléotide le plus exceptionnellement rare dans *M2*, il est tout de même en 7ème position (Tableau 6.2).

Le passage de *M1* à *M2* s'est traduit par le fait que l'espérance s'est rapprochée du comptage, ce qui est normal puisque l'on s'ajuste mieux à la séquence, mais il n'en reste pas moins que la fréquence de CTAG reste inattendue dans *M2*.

- GAT et ATC sont parmi les six tri-nucléotides les plus exceptionnellement fréquents dans *M1* (Tableau 6.3). Lorsque l'on regarde maintenant GATC dans *M1*, il n'est que le 58ème mot le plus fréquent, donc ne fait pas partie des mots les plus exceptionnels (Tableau 6.4). Par ailleurs, GATC est un mot exceptionnellement rare dans le modèle *M2* (Tableau 6.2). Ainsi, sa présence parmi les mots plutôt fréquents dans *M1* est seulement due à ses deux sous-mots GAT et ATC sur-représentés (Figure 6.2). En effet, lorsque l'on tient compte de la sur-représentation de GAT et ATC (modèle *M2*), le nombre d'occurrences de GATC est beaucoup plus grand qu'il n'est prédit et GATC devient sous-représenté dans *M2*.

Tab. 6.3 – Statistiques des six tri-nucléotides les plus fréquents dans M1

W	$N(W)$	$n\widehat{\mu}_1(W)$	$\widehat{\sigma}_1^2(W)$	$U_1(W)$	Rang
CAG	11148	7233.046	4098.548	61.1521	1
CTG	11450	7551.063	4200.194	60.1605	2
ACC	7862	6120.480	3724.538	28.5359	3
GGT	8481	6688.294	4008.440	28.3153	4
ATC	9249	7531.471	4237.498	26.3845	5
GAT	9423	7834.563	4347.907	24.0896	6

Tab. 6.4 – Statistique de GATC dans M1

W	$N(W)$	$n\widehat{\mu}_1(W)$	$\widehat{\sigma}_1^2(W)$	$U_1(W)$	Rang
GATC	2163	1857.749	1569.950	7.7040	58

- GGC et GCC sont deux tri-nucléotides plutôt fréquents dans *M1*, alors que GGCC devient très rare, toujours dans *M1* (Tableau 6.5). Cela signifie que la séquence a tendance à avoir beaucoup de GGC et GCC, mais pas l'un à côté de l'autre. Si l'on s'ajuste maintenant sur la fréquence des tri-nucléotides, en particulier celle de GGC et GCC, GGCC devient le tétra-nucléotide le plus exceptionnellement rare (Tableau 6.2).

Tab. 6.5 – Statistiques de GGC, GCC et GGCC dans M1

W	$N(W)$	$n\widehat{\mu}_1(W)$	$\widehat{\sigma}_1^2(W)$	$U_1(W)$	Rang [1]
GCC	10157	9628.107	4954.966	7.5136	26
GGC	10423	9918.597	5148.273	7.0299	27
GGCC	1478	2257.673	1874.072	-18.0102	12

- Voici l'exemple d'un mot qui est exceptionnel dans *M1*, uniquement parce que l'un de ses sous-mots l'est aussi : c'est le phénomène de contamination par les sous-mots. CAGC est un des tétra-nucléotides les plus exceptionnellement fréquents dans *M1* (Tableau 6.6). Or, lorsque que l'on regarde la pyramide de CAGC dans *M1* (Figure 6.2), on s'aperçoit qu'un de ses sous-mots, CAG, est très fréquent (c'est d'ailleurs le plus fréquent, Tableau 6.3).

1. Le rang indiqué ici est soit le rang de W parmi les mots les plus sur-représentés si la statistique $U_1(W)$ est positive, soit le rang de W parmi les mots les plus sous-représentés si la statistique est négative.

TAB. 6.6 – STATISTIQUE DE CAGC DANS M1

W	$N(W)$	$n\widehat{\mu}_1(W)$	$\widehat{\sigma}_1^2(W)$	$U_1(W)$	Rang
CAGC	3933	2328.712	1916.883	36.6424	3

Par contre, lorsque l'on se place dans *M2*, CAGC n'est plus du tout exceptionnel (Tableau 6.7 et Figure 6.1).

TAB. 6.7 – STATISTIQUE DE CAGC DANS M2

W	$N(W)$	$n\widehat{\mu}_2(W)$	$\widehat{\sigma}_2^2(W)$	$U_2(W)$	Rang
CAGC	3933	3810.561	1439.744	3.2268	78

- À l'inverse, un mot comme GTAG est très bien prédit dans le modèle *M1* et devient exceptionnellement fréquent dans le modèle *M2* (Figure 6.1). En fait, si l'on regarde sa pyramide dans les modèles maximaux (Figure 6.2), on remarque que GTA est fréquent dans *M1*, alors que TAG est très rare dans *M1* ; ces deux effets opposés semblent se compenser de sorte que GTAG n'est plus exceptionnel dans *M1*. D'où l'intérêt de se placer dans le modèle *M2* pour passer outre ces compensations éventuelles.

Ainsi, pour rechercher des mots ayant une fréquence exceptionnelle dans une séquence par rapport à un modèle, il y a deux démarches.

– La première consiste à étudier un mot donné en parcourant l'ensemble des modèles possibles, pour éventuellement détecter une rupture lors du passage d'un modèle à un autre (par exemple, GTAG devient exceptionnel dans *M2* et CAGC perd son caractère exceptionnel dans *M2*).

– La deuxième démarche consiste plutôt à regarder un mot donné dans le modèle d'ordre maximal afin d'éliminer, d'une part, les phénomènes que l'on pouvait déjà observer à partir des mots plus petits (contamination), pour n'observer que des choses nouvelles, et d'autre part, les compensations entre sous-mots qui au contraire peuvent masquer des phénomènes intéressants.

En conclusion, l'étude systématique dans chacun des modèles possibles apporte une information complète sur le caractère exceptionnel d'un mot dans une séquence ; c'est pourquoi, dans les applications traitées par la suite, on se placera successivement dans chaque modèle.

6.4 Choix de l'approximation

On a vu théoriquement que l'approximation gaussienne (Chapitre 2) et l'approximation par une loi de Poisson composée (Partie II) du nombre $N(W)$ d'occurrences d'un mot W dans une séquence sont valides dans l'un des deux cadres asymptotiques respectifs : $n\mu_m(W) \to +\infty$ et $n\mu_m(W)$ =O(1) lorsque n tend vers l'infini (m étant l'ordre du modèle). En pratique, le choix de l'approximation se pose car n est fixé par la longueur de la séquence étudiée ainsi que la longueur h du ou des mot d'intérêt.
On peut cependant dégager quelques règles de décision, illustrées par les graphes des densités de la loi Normale et de la loi de Poisson composée limites associées dans différent cas ; pour cela, on a utilisé une séquence de *E. coli* de longueur 111402 bp.

- Si $n\mu_m(W)$ est trop proche de zéro, l'approximation gaussienne est mauvaise ; on choisira la loi de Poisson composée (Figure 6.3).

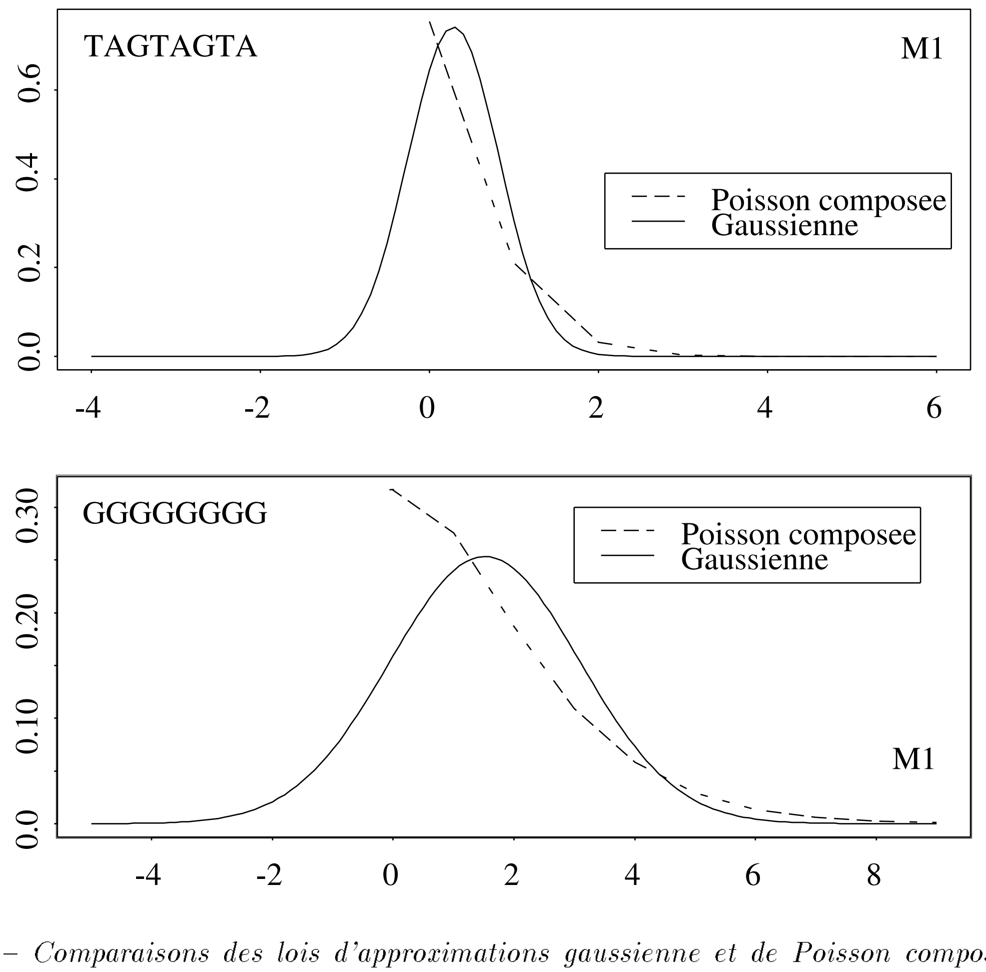


FIG. 6.3 – *Comparaisons des lois d'approximations gaussienne et de Poisson composée pour les mots TAGTAGTA et GGGGGGGG, dont les comptages estimés sous M1 sont respectivement 0.284 et 1.515.*

- Si $(h-m)$ est “petit”, ce qui est typiquement le cas lorsque l'on se place dans le modèle d'ordre maximal $m = h - 2$, l'approximation par la loi de Poisson composée n'est pas conseillée, car elle nécessite tout de même que $(h-m)$ soit asymptotiquement au moins un $\mathrm{O}(\log n)$ (Figure 6.4).
- De plus, si le mot W n'est pas ou “peu” chevauchant, ce qui signifie que ses périodes principales, si elles existent, sont supérieures à $(h-m)$, la loi de Poisson composée se réduit à une loi de Poisson (Chapitre 5). Par conséquent, l'approximation de Poisson sera mauvaise dès lors que l'espérance et la variance de $N(W)$ sont différentes. On a constaté que lorsqu'on augmente la longueur des mots, m étant fixé, ou inversement lorsque l'on diminue l'ordre du modèle pour un mot fixé, l'écart entre l'espérance et la variance des mots peu chevauchants décroît vers zéro ; cela rejoint l'idée du point précédent. Si l'on reprend l'exemple de TGGGTG (Figure 6.4), les valeurs estimées de l'espérance et la variance du comptage sont : (33.633,32.958) dans *M1* , (35.412,33.546) dans *M2* , (31.687,25.673) dans *M3* et (33.095,14.876) dans *M4*.

Cependant, les deux approximations seront quasi équivalentes dans un grand nombre de cas (Figure 6.4, modèles *M1* et *M2*).

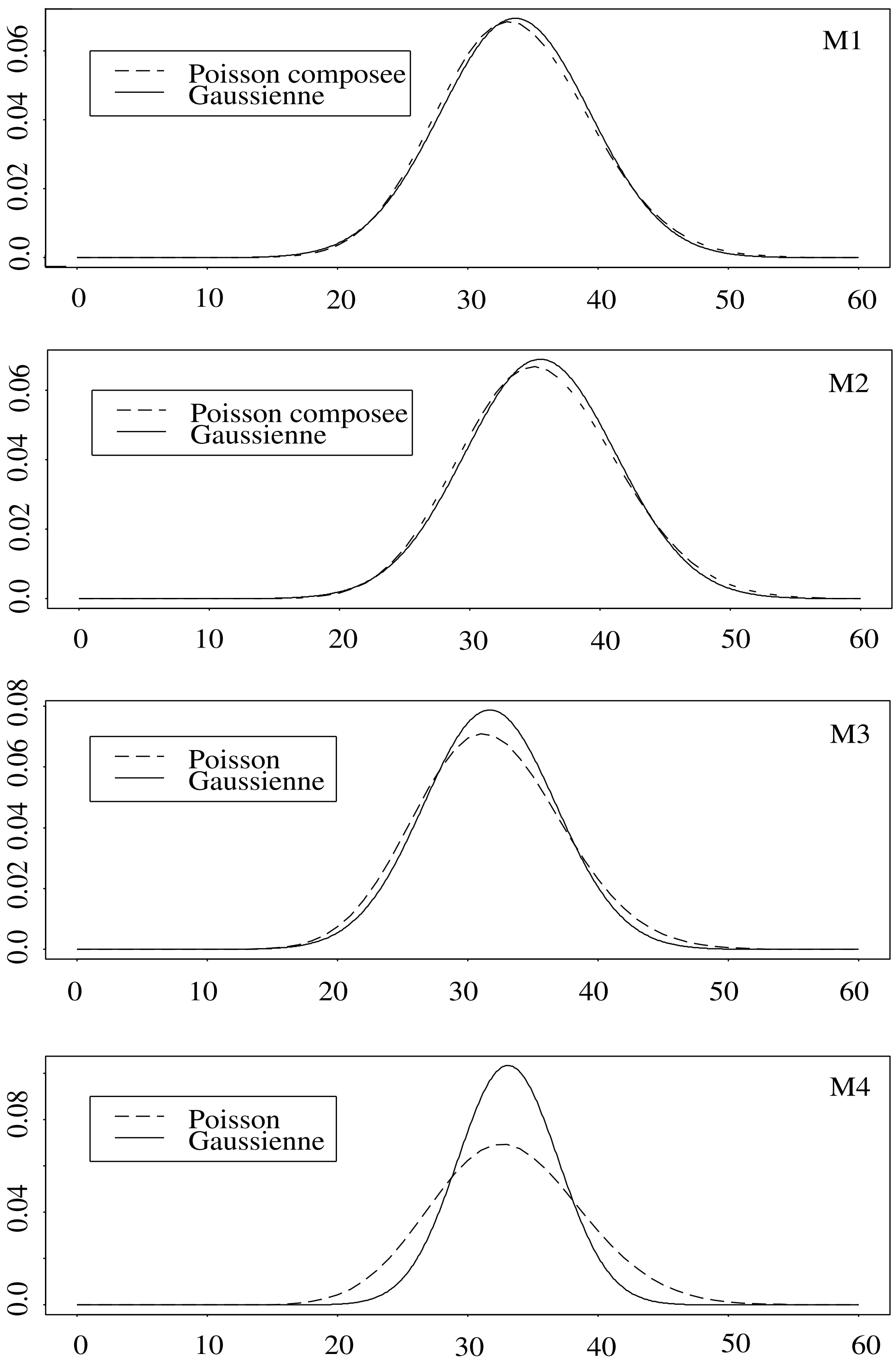


FIG. 6.4 – *Comparaisons des lois d'approximations gaussienne et de Poisson composée pour le mot TGGGTG sous les modèles M1 à M4. La loi de Poisson composée se réduit à une loi de Poisson dans M3 et M4.*

Le choix de l'approximation semble toutefois moins important pour la recherche de mots de fréquence exceptionnelle. En effet, si les calculs approchés de la queue de distribution de $N(W)$ diffèrent quelque peu en utilisant l'une ou l'autre des deux approximations, le classement des mots de fréquence exceptionnelle ne semble pas être affecté (Figure 6.5).

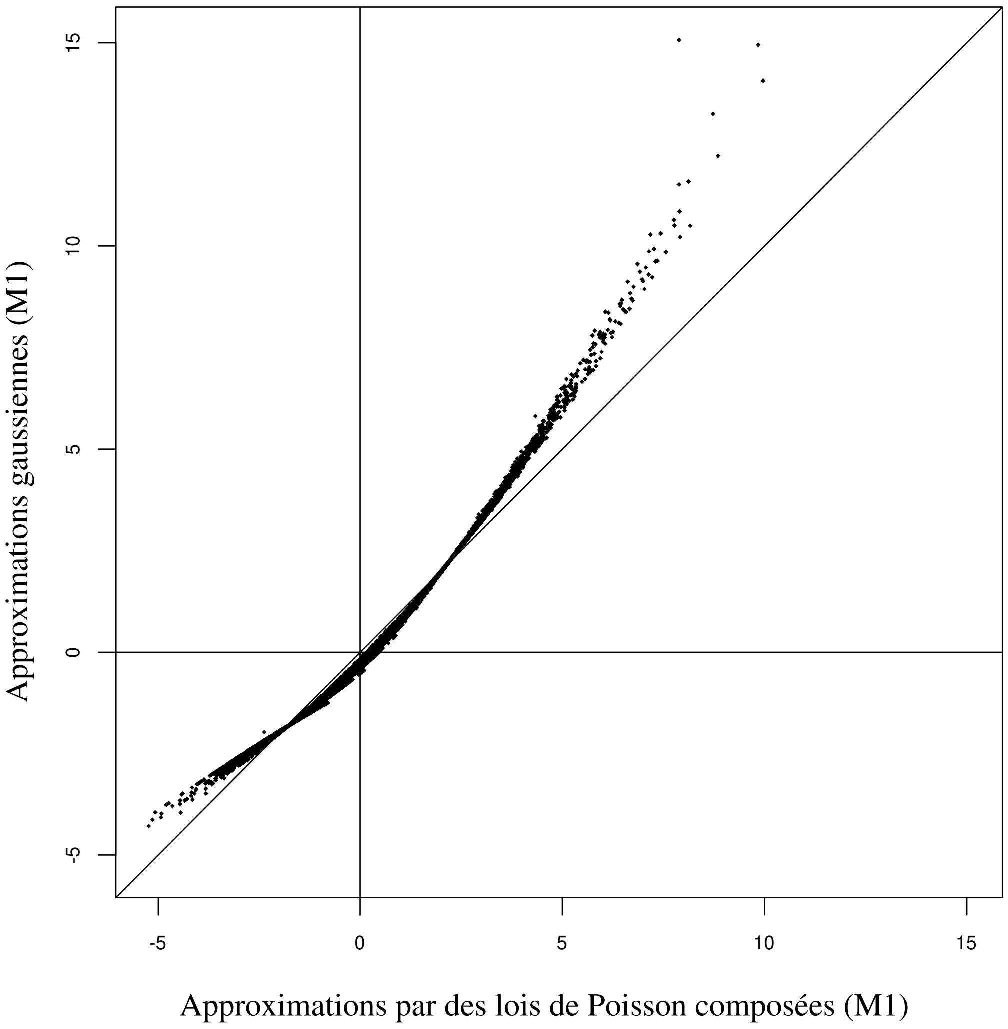


FIG. 6.5 – *Comparaison de l'approximation gaussienne et de l'approximation par une loi de Poisson composée pour la recherche de 7-mots exceptionnels dans M1.*

Chaque 7-mot est représenté par un point dont les coordonnées sont les quantiles de la loi Normale centrée réduite associés aux P-valeurs calculées à partir de l'approximation gaussienne et de Poisson composée, dans le modèle M1.

Calculs de grandes déviations

Afin de comparer les queues de distribution obtenues par les deux approximations pour des mots exceptionnels, nous avons calculé celle obtenue par les méthodes de grandes déviations dans les chaînes de Markov (Annexe D). Dans le Tableau 6.8, nous avons reporté les quantiles de la loi Normale centrée réduite associés aux trois queues de distributions (approximation gaussienne, approximation de Poisson composée, approche grandes déviations), pour chacun des dix 4-mots les plus exceptionnellement rares et des dix 4-mots les plus exceptionnellement fréquents obtenus par l'approximation gaussienne. Nous nous sommes placés dans *M1* et la séquence étudiée est de longueur 111402.

TAB. 6.8 – COMPARAISON DES QUANTILES

mot	comptage observé	comptage estimé	variance estimée	quantile [G]	quantile [PC]	quantile [GD]
CTTG	204	506.146	435.016	-14.4865	-15.2425	-15.2123
CTAG	21	208.947	187.227	-13.7357	-16.6318	-16.5444
CAAG	214	458.256	396.794	-12.2620	-12.7118	-12.6033
CATG	360	605.180	485.950	-11.1222	-10.7564	-10.6397
TGTG	336	570.540	466.452	-10.8596	-10.0076	-9.8928
GAGG	180	357.121	307.659	-10.0980	-10.2315	-10.0390
CCTA	78	213.376	191.317	-9.7873	-10.6048	-10.3986
CGAG	253	432.632	364.448	-9.4095	-9.3329	-9.1177
GGCC	336	532.200	439.256	-9.3614	-9.0951	-8.9029
TAGG	104	240.208	213.665	-9.3183	-9.8598	-9.6337
CGGT	721	466.642	408.218	12.5893	10.9203	10.7714
TCAG	562	344.605	296.749	12.6199	10.7527	10.5460
TATC	489	279.735	244.507	13.3829	11.3339	11.1256
CAGG	622	380.072	325.341	13.4127	11.3831	11.1969
TGGT	708	438.815	385.792	13.7048	(8.5670)	11.4412
GGCG	1110	764.447	601.086	14.0944	(7.8823)	11.3878
CAGC	851	504.457	416.507	16.9803	(8.0341)	13.7318
GCTG	1073	625.336	498.605	20.0482	(7.8592)	15.9753
CCAG	736	348.802	300.009	22.3546	18.0511	18.0472
CTGG	962	432.957	365.429	27.6751	21.8884	22.0442

On remarque, dans l'ensemble, que le quantile obtenu par l'approche grandes déviations est plus proche de celui obtenu par la loi de Poisson composée que de celui obtenu par l'approximation gaussienne. L'approximation gaussienne semble sous-estimer les mots exceptionnellement rares, et au contraire sur-estimer les mots exceptionnellement fréquents. Cependant, les mots les plus exceptionnels pour les deux approximations sont les mêmes, à l'ordre près. Les quantiles entre parenthèses sont à prendre avec précautions car ils semblent liés à une erreur numérique lors du calcul de la queue de distribution de la loi de Poisson composée associée (un quantile de l'ordre de 8 correspond à une queue de distribution de l'ordre de 10^{-16}, c'est à dire la précision de l'ordinateur). Cela est dû à un comptage N trop élevé qui rend difficile le calcul de la probabilité pour que la loi de Poisson composée ad-hoc dépasse N. En effet, cela nécessite de prendre en compte toutes les partitions de l'entier N, c'est à dire toutes les possibilités d'écrire N sous la forme $\sum_{k>0} k\, N_k$ avec $N_k \in \mathbb{N}$. À priori, l'approximation de Poisson composée est adaptée pour des mots rares.

6.5 Boîte à outils pour la recherche de mots exceptionnels

Nous avons créé une boîte à outils informatique, destinée à être diffusée au sein de la communauté des Biologistes, qui permette de rechercher de façon automatique les mots de fréquence exceptionnelle dans une séquence d'ADN donnée (Gélis et Schbath, 1995). La boîte à outils "Recherche de Mots Exceptionnels dans une Séquence" (RMES) contient deux types de programmes.

- Les programmes de calculs proprement dit, mettent en oeuvre les différentes méthodes d'approximations décrites dans les deux premières parties. Ils calculent, pour chaque mot d'une longueur donnée, et pour un ordre de modèle donné, une statistique en échelle gaussienne standard qui indique le caractère exceptionnel du mot associé, comme cela est décrit au paragraphe 6.2.
  Pour l'approximation gaussienne, trois approches ont été exploitées : l'approche martingale, l'approche conditionnelle et l'approche conditionnelle asymptotique, chacune associée à la façon de calculer les estimateurs de l'espérance et de la variance du comptage (Partie I). Ces différentes approches ont été développées aussi bien dans les modèles sans phase, qu'avec phase.
  Pour les approximations de Poisson, deux programmes ont été créés: l'un implémente l'approximation du nombre de trains d'un mot par une loi de Poisson, et l'autre, l'approximation par une loi de Poisson composée du nombre d'occurrences d'un mot.
  Cet ensemble de programmes a été développé en langage C++.

- Des programmes de visualisations des résultats décrits ci-dessus, composent la deuxième partie de RMES. Mis à part un programme qui liste les mots par ordre croissant de leur statistique sous forme de tableaux (Tableau 6.2, par exemple), les résultats sont visualisés de façon graphique par l'intermédiaire de fonctions développées en Splus. Ces représentations graphiques sont essentiellement de deux sortes :

– La représentation globale, sous forme de nuage de points (voir Figure 6.1), est un bon outil pour comparer l'effet de deux modèles ou comparer deux séquences.

– La représentation sous forme de pyramide permet, quant à elle, de visualiser simultanément le caractère exceptionnel d'un mot et de ses sous-mots, soit pour un modèle donné, soit dans le modèle maximal associé à chaque longueur de mot (voir Figure 6.2).

Chapitre 7

Étude des 3- et 4-mots phasés dans les génomes de *E. coli* et *B. subtilis*

Cette étude utilise l'approximation gaussienne de la loi du nombre d'occurrences d'un mot dans une séquence.

7.1 Les 3-mots phasés dans *E. coli*

Nous avons ajusté les modèles *M1_3* et *M2_3* sur un ensemble de 184 gènes de *Escherichia coli* (181083 bp) extraits de quatre séquences (L10328, M87049, L19201 et U00006 dans Genbank) et situés à la fois sur le même brin d'ADN et du même côté de l'origine de réplication. Les gènes codant sur le brin complémentaire ne sont donc pas pris en compte.

• **Les 3-mots exceptionnels sous** *M1_3*

Le calcul des statistiques des 3-mots phasés dans le modèle d'ordre 1 montre que les mots sont principalement exceptionnels sur la phase 3 (les codons). Pour les trois différentes phases, l'éventail des statistiques est :

$$\begin{array}{rcccr} -12 & < & U_{1-3}(abc,1) & < & 9\,, \\ -13 & < & U_{1-3}(abc,2) & < & 19\,, \\ -30 & < & U_{1-3}(abc,3) & < & 48\,. \end{array}$$

Une importante partie de l'écart entre le modèle et les comptages des 3-mots phasés est ainsi expliquée par la fréquence des codons. Aussi, la Figure 7.1 où l'on a représenté les statistiques globales (toutes phases confondues) $U_{1-3}(abc)$ contre les statistiques sur la phase 3, $U_{1-3}(abc,3)$, illustre clairement cette remarque. En effet, très peu de points s'éloignent de la diagonale $x = y$.

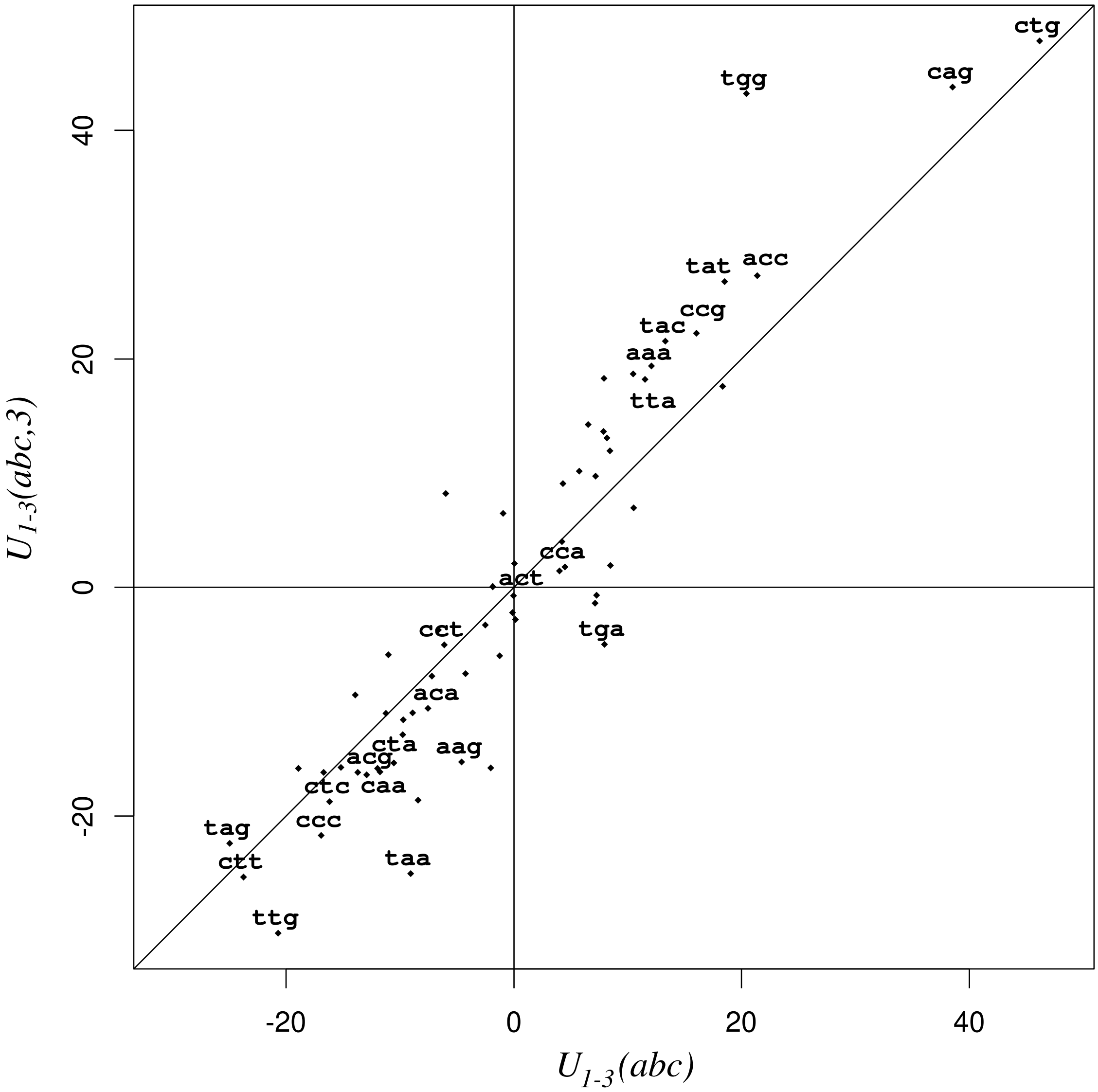


FIG. 7.1 – *Exceptionnalité des 3-mots dans E. coli.*

Comme la statistique $U_{1-3}(abc)$ s'obtient à partir de la différence entre le comptage global $N(abc)$ et son estimation, l'axe des abscisses résume l'écart au modèle des comptages d'un mot sur les trois phases. L'axe des ordonnées montre uniquement l'écart au modèle des comptages sur la phase 3 (phase de lecture).

Le Tableau 7.1 donne la liste des mots abc possédant une statistique sur la phase 3 $|U_{1-3}(abc,3)| > 19$. Certains d'entre eux sont également exceptionnels sur la phase 2 avec une statistique $|U_{1-3}(abc,2)| > 10$, et TAG l'est de plus sur la phase 1.

Tab. 7.1 – Les 3-mots exceptionnels dans *E. coli.*

abc	$U_{1-3}(abc,3)$	$U_{1-3}(abc,2)$	*abc*	$U_{1-3}(abc,3)$	$U_{1-3}(abc,2)$	$U_{1-3}(abc,1)$
<u>CTG</u>	48	16	<u>TTG</u>	-30	-11	
<u>CAG</u>	44	19	<u>CTT</u>	-25	-13	
<u>TGG</u>	43		TAA	-25		
<u>ACC</u>	27		<u>TAG</u>	-22	-10	-10
<u>TAT</u>	27		<u>CCC</u>	-22		
<u>CCG</u>	22					
TAC	22					
AAA	19					

Seuls les mots abc ayant une statistique $|U_{1-3}(abc,3))| > 19$ *sont mentionnés. Leurs statistiques sur les phases 1 et 2 sont indiquées si elles sont supérieures à 10 en valeurs absolues. Les mots soulignés ont aussi été détecté par Brendel* et al. *(1986).*

En reprenant les codons exceptionnels du Tableau 7.1 et en examinant les statistiques de leurs codons synonymes, c'est à dire des codons qui codent pour le même acide aminé, données dans le Tableau 7.2, on remarque que l'unique codon TGG qui code pour l'acide aminé tryptophane est sur-représenté. Il en est de même pour les deux codons de tyrosine, TAT et TAC. Toutefois, les autres codons du Tableau 7.1 possèdent des codons synonymes ayant une grande statistique de signe opposé. Par conséquent, l'écart au modèle *M1_3* ne s'explique pas par la composition en acides aminés de la protéine codée, mais par l'usage des codons.

Tab. 7.2 – Codons exceptionnels et leurs codons synonymes dans *E. coli.*

abc ($U_{1-3}(abc,3)$)	acide aminé
CTG (48) TTA (18) CTA (-16) CTC (-19) CTT (-25) TTG (-30)	leucine
CAG (44) CAA (-16)	glutamine
TGG (43)	tryptophane
ACC (27) ACT (-1) ACA (-11) ACG (-16)	threonine
TAT (27) TAC (22)	tyrosine
CCG (22) CCA (2) CCT (-5) CCC (-22)	proline
AAA (19) AAG (-15)	lysine
TAA (-25) TAG (-22) TGA (-5)	stop

Chaque codon du Tableau 7.1 ainsi que leurs synonymes sont indiqués ici. Les codons les plus sous-représentés apparaissent tous comme synonyme d'un codon sur-représenté, sauf pour les trois codons stop. Les mots de ce tableau sont identifiés par leur nom sur la Figure 7.1.

• **Les 4-mots exceptionnels sous** *M1_3*

Les 4-mots exceptionnels sous *M1_3* apparaissent principalement sur les phases 1 ou 3, comme le montre l'éventail des statistiques sur les différentes phases :

$$\begin{array}{rcccl} -20 & < & U_{1-3}(abcd,1) & < & 34\,, \\ -9 & < & U_{1-3}(abcd,2) & < & 15\,, \\ -22 & < & U_{1-3}(abcd,3) & < & 43\,. \end{array}$$

Ceci s'explique simplement par les deux faits suivant :

– les 4-mots en phase 1 ou 3 possèdent un sous-mot de longueur 3 en phase 3, autrement dit un codon, tandis que ceux en phase 2 n'en possèdent pas,

– les 4-mots exceptionnels ici contiennent au moins un sous-mot de longueur 3 exceptionnel dans le même modèle.

Ce deuxième point apparaît dans le Tableau 7.3 où figurent, pour chacune des trois phases, les 4-mots ayant une statistique $|U_{1-3}(abcd,k)| > 15$. On s'aperçoit en effet que presque tous ces 4-mots possèdent un sous-mot exceptionnel soit sur la phase 3 (entre crochet), soit sur la phase 2 (entre parenthèse).

TAB. 7.3 – LES 4-MOTS EXCEPTIONNELS SOUS M1_3 DANS *E. coli*.

Phase 1		Phase 2		Phase 3	
abcd	$U_{1-3}(abcd,1)$	*abcd*	$U_{1-3}(abcd,2)$	*abcd*	$U_{1-3}(abcd,3)$
∗[CTG]G	34	C(CAG)	15	∗C[TGG]	44
[CAG]G	30			G[CTG]	30
[TGG]C	24			(CAG)C	25
▷[TAT]C	22			T[CAG]	24
[TGG]T	20			C[CAG]	21
[TGG]G	18			G[TGG]	20
[AAA]G	17			G[CAG]	19
[CTG]A	17			▷TATC	17
[CAG]C	17			C[ACC]	17
GAAG	16			C[TAC]	16
				T[TAT]	16
				GGCG	15
[TTG]G	-20			C[TTG]	-22
CAAG	-16				
[TAA]G	-15				

Pour chacune des trois phases, les 4-mots possédant une statistique $|U_{1-3}(abcd,k)| >$ 15 *sont listés. Leurs sous-mots de longueur 3 exceptionnels sur la phase 3 (codons)* *sont mis entre crochets, tandis que ceux sur la phase 2 sont entre parenthèses. À l'exception de CCG, tous les codons sur-représentés du Tableau 7.1 possèdent au moins une extension de longueur 4 exceptionnelle. On remarque qu'il y a peu de 4-mots sous-représentés. Deux mots apparaissent exceptionnels sur deux phases différentes : ils sont marqués avec un signe $*$ ou $\triangleright$.*

Aussi, cette liste de 4-mots exceptionnels nous apporte une information supplémentaire à celle des 3-mots, à savoir que seules certaines extensions de 3-mots exceptionnels sont des 4-mots exceptionnels.

• Les 4-mots exceptionnels sous *M2_3*

Étant donné que le modèle *M2_3* prend en compte les comptages des 3-mots, on s'attend à obtenir des 4-mots exceptionnels différents de ceux obtenus précédemment. Le Tableau 7.4 nous permet de constater les deux choses suivantes :

– certains 4-mots sont exceptionnels à la fois dans les modèles *M1_3* et *M2_3*,

– si l'on change une lettre des mots sur-représentés suivants : CAGG, TTGC, CTGG, GGCG et CGCC, on obtient les mots sous-représentés CAAG, TTGG, TTGG, GGCC et CGCG.

• Les gènes du brin complémentaire

La même analyse appliquée aux gènes du brin complémentaire montre que les mêmes mots sont exceptionnels avec des statistiques semblables (non présenté). Cela semble signifier que l'information portée par les 3- et 4-mots est principalement liée à la fonction codante des séquences et non à la structure linéaire d'un brin.

• Comparaison avec Brendel *et al.* (1986)

Brendel *et al.* (1986) donnent une liste de mots exceptionnels dans quatre séquences composées de gènes de *E. coli*. Pour cela, ils utilisent le modèle *M1* pour les 3-mots, et le modèle *M2* pour les 4-mots. Nos résultats peuvent être comparés aux leurs dans la mesure où les 3-mots exceptionnels sous *M1_3* et *M1* sont les mêmes et dans le même ordre, dans *E. coli* (non présenté), et d'autre part, où on a vu que les 3- et 4-mots exceptionnels étaient identiques pour des gènes codant sur un brin ou sur l'autre.

Beaucoup de mots sont communs aux deux études (ils sont soulignés dans les Tableaux 7.1 et 7.4), néanmoins des différences subsistent. Brendel *et al.* (1986) trouvent des mots exceptionnels que nous n'avons pas : GAT(+), GGT(+), TGA(+), GAG(-), GTC(-),
GGG(-) et AATA(+), CTAC(+), GAAG(+), GATG(+), TGCT(+), TTCC(+), CTAG(-), GAAT(-), GGCT(-), GGTT(-), GTCC(-),
et nous trouvons par contre TAC(+), AAA(+), AGC(+), GTA(+), TAA(-), ATA(-), et TTGC(+), ACGC(+), ATTC(+), CTTC(+), TGCC(+), AGAG(+), TATC(+), GCAA(+),

CGAC(+), GAGG(-), CTTG(-), CGTG(-), CGCG(-).
Pour les mots chevauchants, la différence peut être expliquée par les occurrences chevauchantes de ces mots qui doivent être prises en compte dans l'estimation de la variance. Par exemple, pour AAA nous estimons l'espérance du comptage par 1562 et la variance par 771, et pour GGG nous avons respectivement 786 et 383, tandis que Brendel *et al.* (1986), en faisant une approximation par une loi de Poisson, considèrent la variance égale à l'espérance.

Notre étude permet de mettre en évidence, non seulement que les 3-mots sont exceptionnels sur la phase 3, la phase des codons, mais aussi que les 4-mots sont principalement exceptionnels sur une des phases 1 ou 3, phases correspondant à des 4-mots qui commencent ou terminent par un codon.

TAB. 7.4 – LES 4-MOTS EXCEPTIONNELS SOUS M2_3 DANS *E. coli*.

Phase 1		Phase 3	
abcd	$U_{2-3}(abcd,1)$	*abcd*	$U_{2-3}(abcd,3)$
<u>CAGG</u>	15	<u>GGCG</u>	11
TTGC	10	CTTC	10
ACGC	10	TGCC	10
ATTC	10	AGAG	9
<u>CTGG</u>	9	<u>CGCC</u>	9
<u>TTCC</u>	9	**TATC**	9
<u>CAAC</u>	8	<u>GGTG</u>	8
		GCAA	8
		CGAC	8
<u>CAAG</u>	-16	∗<u>GGCC</u>	-17
•**<u>TTGG</u>**	-12	<u>CCAA</u>	-10
GAGG	-11	**CTTG**	-10
∗<u>GGCC</u>	-10	•<u>TTGG</u>	-10
<u>TAAG</u>	-8	CGTG	-8
		CGCG	-8

Pour chacune des phases, on liste les 4-mots qui ont une statistique $|U_{2-3}(abcd,k)| > 8$. *Les mots apparaissant en gras sont exceptionnels à la fois dans M1_3 et M2_3. Ceux qui sont exceptionnels sur plusieurs phases sont marqués d'un même signe (∗, •). Les mots soulignés ont aussi été détecté par Brendel* et al. *(1986).*

7.2 Les 3-mots phasés dans *B. subtilis*

Comme au paragraphe précédent, nous avons ajusté les modèles *M1_3* et *M2_3* sur un ensemble de 159 gènes de *Bacillus subtilis* (134559 bp) extraits de deux séquences (X73124 dans Genbank et A. Sorokine, communication personnelle). Ces gènes sont situés sur le même brin d'ADN et du même côté de l'origine de réplication.

• Les 3-mots exceptionnels sous *M1_3*

Les 3-mots exceptionnels sont de nouveau principalement les codons mais, contrairement à l'étude précédente faite sur *E. coli*, il y en a aussi sur la phase 2. L'éventail des statistiques sur chacune des phases est :

$$\begin{aligned} -10 &< U_{1-3}(abc,1) < 12\,, \\ -18 &< U_{1-3}(abc,2) < 26\,, \\ -27 &< U_{1-3}(abc,3) < 34\,. \end{aligned}$$

Le Tableau 7.5 présente les 3-mots ayant une statistique $|U_{1-3}(abc,k))| > 14$ sur au moins une des trois phases. Quatre d'entre eux apparaissent exceptionnels sur deux phases.

TAB. 7.5 – LES 3-MOTS EXCEPTIONNELS DANS *B. subtilis*.

Phase 3		Phase 2	
abc	$U_{1-3}(abc,3)$	*abc*	$U_{1-3}(abc,2)$
TAT	34	CGG	26
→**TGG**	23	∗**CAG**	16
CCG	19	GCT	15
AAA	18		
∗**CAG**	17		
TAC	16		
TTA	15		
ACA	15		
CTG	15		
TAA	-30	→TGG	-18
∘**TAG**	-21	CGA	-16
•AAT	-17	•AAT	-15
TGA	-14	∘**TAG**	-14
CCA	-14		

Les mots en gras sont exceptionnels sur la même phase à la fois dans B. subtilis et E. coli. Les mots exceptionnels sur deux phases différentes sont marqués d'un signe commun (→, ∗, ∘, •).

Remarquons le cas de TGG : la sur-représentation de TGG sur la phase 3 (tryptophane) est compensée par sa sous-représentation sur la phase 2.

Comme dans *E. coli*, la sur-représentation de TGG est liée à la fréquence de l'acide aminé tryptophane, tandis que celle de TAT et TAC est liée à l'acide aminé tyrosine. Les autres codons exceptionnels montrent un biais important dans l'usage des codons. Si nous comparons les 3-mots dans *E. coli* et *B. subtilis*, beaucoup de codons exceptionnels sont communs : ils sont indiqués en gras dans le Tableau 7.5. Deux codons sont par contre spécifiques de *E. coli*, ACC(+27) et CTT(-25), et trois le sont de *B. subtilis*, ACA(+15), TGA (-14) et CCA(-14). Sur la phase 2, CTG(+16) et CTT(-13) sont spécifiques de *E. coli*, et CGG(+26), GCT(+15), TGG(-18) et CGA(-16) le sont de *B. subtilis*.

Nous pouvons résumer la comparaison des 3-mots dans les deux bactéries, en représentant chaque mot par un point de coordonnées égales aux statistiques $U_{1-3}(abc)$ associées avec *E. coli* (abscisses) et *B. subtilis* (ordonnées). Sur la Figure 7.2, les 3-mots exceptionnels communs et spécifiques des deux bactéries sont identifiés avec un indice indiquant la phase sur laquelle ils sont significatifs. Seul TGG a une position particulière qui est due à la combinaison de ses différents comportements sur les phases 2 et 3 (indiqués par des flèches).

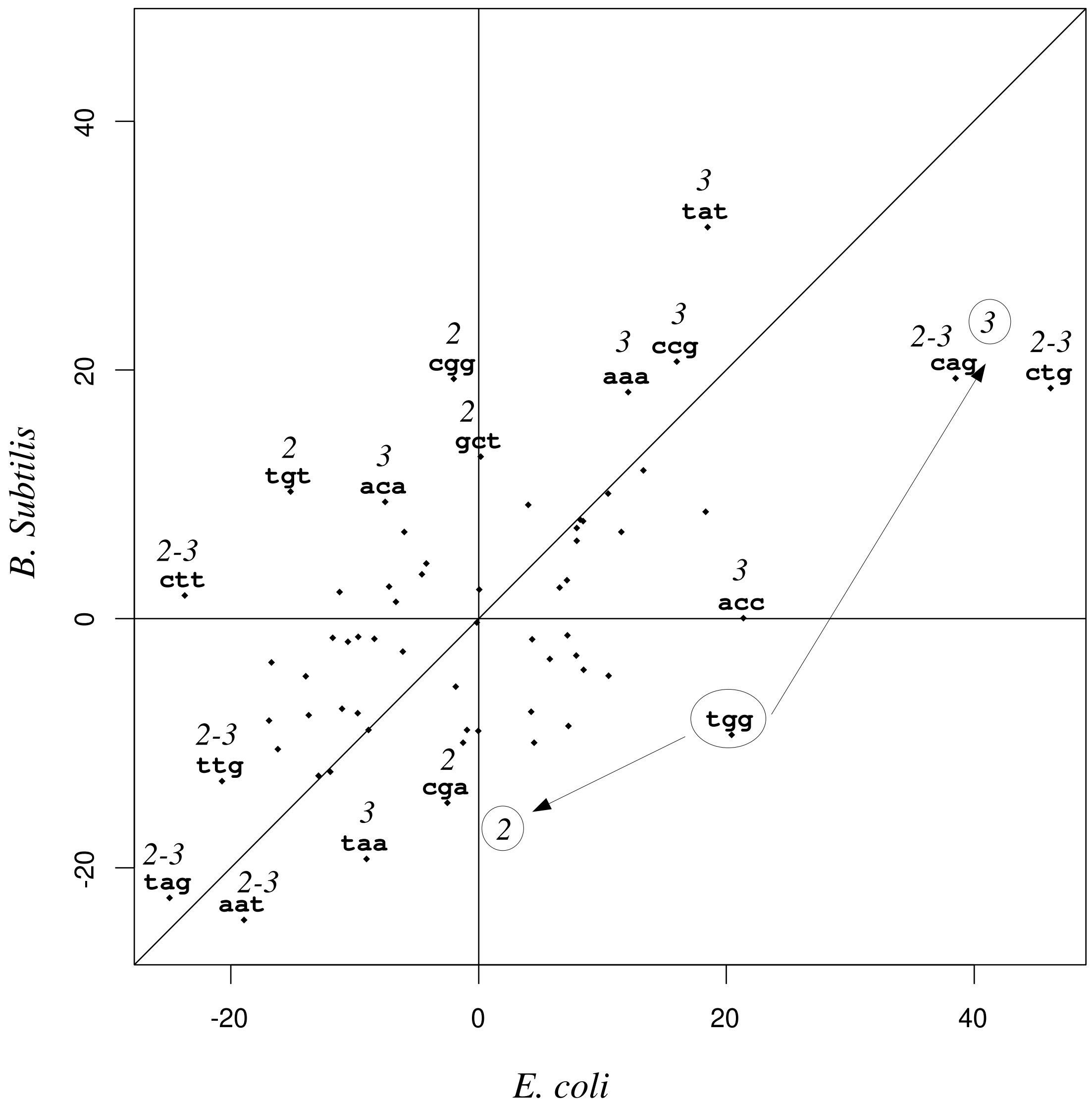


FIG. 7.2 – *Comparaison de E. coli et B. subtilis au travers des 3-mots exceptionnels sous M1_3.*

Chaque 3-mot est représenté avec ses statistiques globales $U_{1-3}(abc)$ dans E. coli contre B. subtilis. Puisque la statistique $U_{1-3}(abc)$ est obtenue à partir du comptage global (toutes phases confondues) $N(abc) = \sum_k N(abc, k)$, cette figure résume l'information sur les trois phases. La position sur ce graphe de tous les mots du bord du nuage correspond à une position très similaire de celle qu'ils auraient dans un ou deux des graphes associés à une phase précise (indiquée au dessus du nom); leurs positions pour les autres phases sont quasi centrales. Seul TGG possède une position sur la figure globale qui ne suit pas cette règle. Les positions de TGG sur les figures des phases 2 et 3 sont montrées par des flèches.

• **Les 4-mots exceptionnels sous** *M2_3*

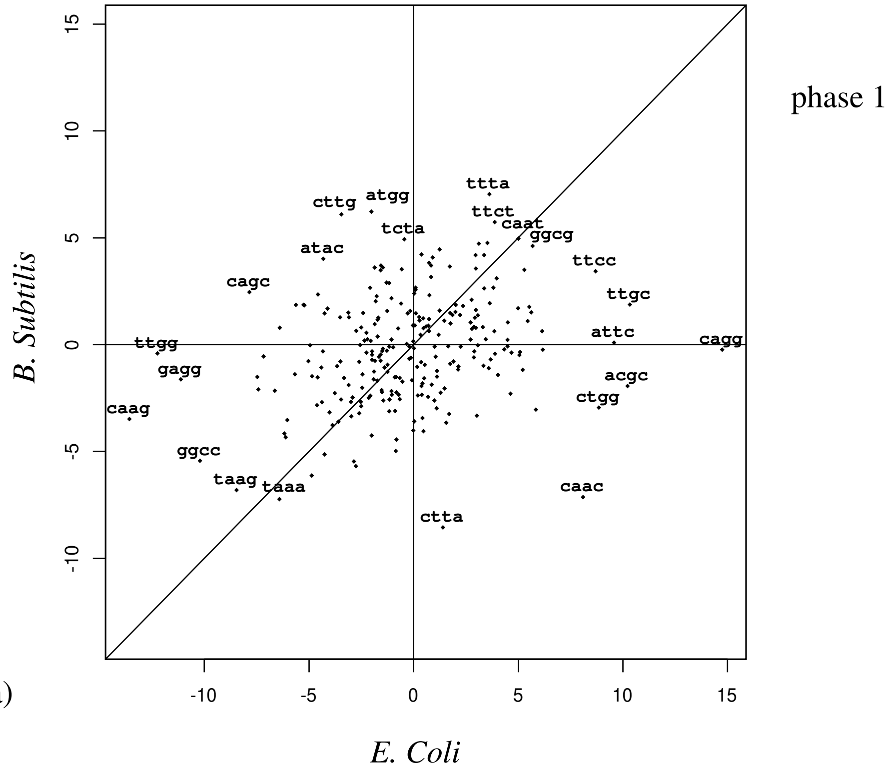
Comme pour *E. coli*, les 4-mots exceptionnels sous *M1_3* sont essentiellement expliqués par des sous-mots exceptionnels de longueur 3 (non montré). C'est pourquoi, nous nous intéressons maintenant aux 4-mots exceptionnels sous *M2_3*.
Le Tableau 7.6 montre les 4-mots exceptionnels chez *B. subtilis* avec une statistique $|U_{2-3}(abcd, k)| > 7$. Seul TAAA est commun aux deux modèles *M1_3* et *M2_3*.

TAB. 7.6 – LES 4-MOTS EXCEPTIONNELS SOUS M2_3 DANS *B. subtilis*.

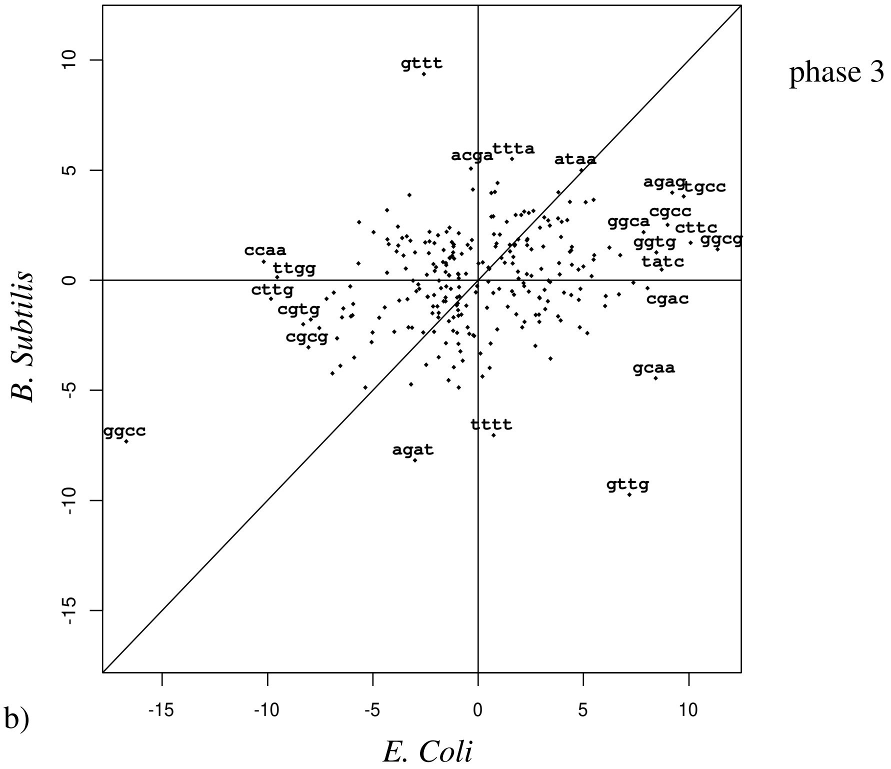
Phase 1		Phase 3	
abcd	$U_{2-3}(abcd, 1)$	*abcd*	$U_{2-3}(abcd, 3)$
TTTA	7	GTTT	9
CTTA	-8	GTTG	-9
TAAA	-7	AGAT	-8
CAAC	-7	GGCC	-7
		TTTT	-7

La comparaison des deux bactéries révèle que très peu de 4-mots exceptionnels sont communs. Ceci est aussi illustré par la Figure 7.3 où les nuages de points correspondant aux phases 2 et 3 ont une forme assez ronde.

Figure 7.3 : Les 4-mots sont représentés à l'aide de leur statistiques $U_{2-3}(abcd, 1)$ dans a), et $U_{2-3}(abcd, 3)$ dans b), dans E. coli contre B. subtilis. Les mots du bord des nuages sont nommés, tandis que les autres sont simplement représentés par un point. À l'exception de quelques mots comme GGCC en phase 3 et TTTA en phase 1, la plupart des 4-mots exceptionnels sont spécifiques d'une des deux bactéries.



a)



b)

FIG. 7.3 – *Comparaison de E. coli et B. subtilis au travers des 4-mots exceptionnels sous M2_3.*

Chapitre 8

Étude des palindromes dans *E. coli* et les phages *Lambda* et *T7*

La structure double brin de l'ADN nous conduit à définir le conjugué d'un mot W. Le conjugué de W, noté $\overline{W}$, est le mot correspondant à W sur l'autre brin. Par conséquent, il s'obtient en renversant le mot W puis en remplaçant chacune des lettres par sa lettre complémentaire, c'est à dire A$\leftrightarrow$T et C$\leftrightarrow$G. Un palindrome "biologique" est ainsi un motif qui est identique à son conjugué, par exemple ACTAGT ou AGCT.

Le nombre de palindromes est connu pour être significativement faible dans le génome de certains phages, de *E. coli* et de *B. subtilis*. Ce phénomène a été interprété comme un moyen d'éviter les sites "cibles" des enzymes de restriction, car en ces sites les enzymes de restriction coupent le génome. Dans une étude sur les comptages des palindromes dans un large ensemble d'organismes, Karlin *et al.* (1992) ont montré une relation entre ces comptages et l'existence de systèmes de restriction ayant les palindromes pour cible. Différents mécanismes peuvent être envisagés pour éviter ces motifs. Nous approfondissons ici l'étude du caractère exceptionnellement rare des palindromes en considérant leurs statistiques asymptotiquement gaussiennes centrées réduites (Partie I) sous différents modèles.

8.1 Fréquence des 4-, 6-, et 8-palindromes dans *E. coli*

Cette étude a été réalisée sur une séquence contiguë d'ADN de *E. coli* obtenue à partir de quatre séquences (L10328, M87049, L19201 et U00006 dans Genbank) et dont on a conservé uniquement les 384243 paires de bases situées à droite de l'origine de réplication .

• Étude des 4-palindromes sous *M1* et *M2*

La Figure 8.1 montre les statistiques des 4-mots sous les modèles *M1* et *M2* et permet de visualiser la position des palindromes de longueur 4 (identifiés par leur nom).

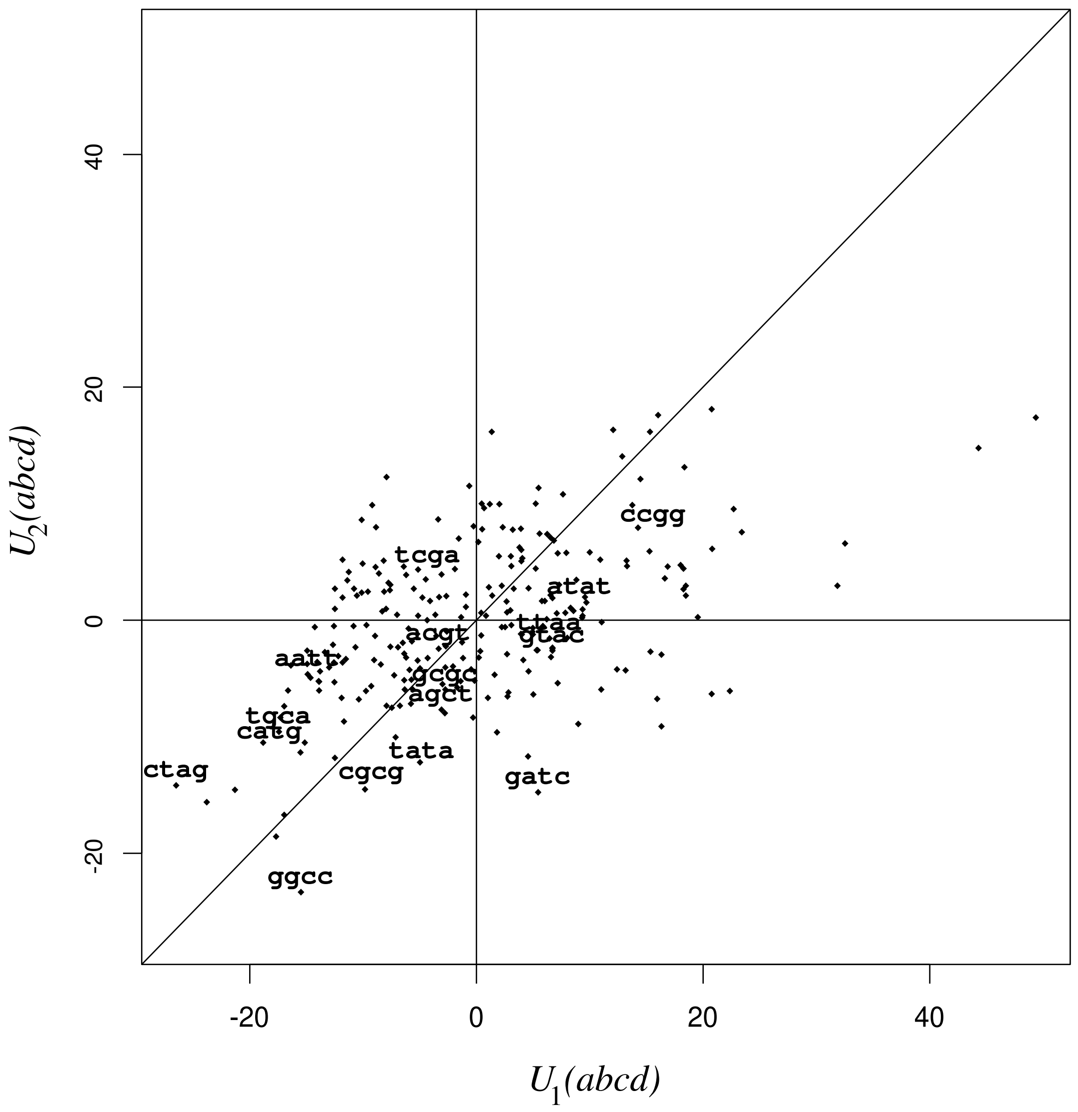


FIG. 8.1 – *Statistiques des 4-mots sous M1 et M2 dans E. coli.*

Chaque mot abcd est représenté par un point de coordonnées $(U_1(abcd), U_2(abcd))$*. Tous les palindromes sont nommés sur la figure.*

La plupart des palindromes sont sous-représentés dans au moins un des modèles. CCGG est le seul palindrome sur-représenté. CTAG est le 4-mot le plus évité lorsque l'on base la prédiction sur les dinucléotides, tandis que GGCC est le plus évité lorsque l'ajustement se fait sur les trinucléotides. Ceci signifie que ces deux mots sont évités par deux voies d'évitement différentes. CTAG est évité principalement parce que CTA et TAG sont eux même évités[1]. Aussi, AATT est sous-représenté dans *M1* uniquement parce que AAT est exceptionnellement rare aussi. En

1. Les statistiques des 3-mots dans *M1* ne sont pas montrées. Cependant, la séquence étudiée est une sous-séquence de celle étudiée en Section 6.3 et si l'on compare les Figures 8.1 et 6.1, on constate que les résultats

effet, lorsque l'on tient compte des fréquences des trinucléotides, AATT n'est plus exceptionnel et CTAG devient le 8ème mot le plus sous-représenté. Au contraire, bien que GGC et GCC soit assez bien représentés, GGCC est exceptionnellement rare car un G est évité devant GCC ou un C est évité après GGC. De plus, GATC est évité en dépit de la sur-représentation de ATC et GAT. Il y a donc une nouvelle information dans l'exceptionnelle rareté de GATC et GGCC qui n'est pas compatible avec la fréquence des trinucléotides.

• **Étude des 6-palindromes sous** *M1* **à** *M4*

Les 6-mots peuvent être analysés dans quatre modèles. La Figure 8.2 montre les statistiques des palindromes comparées aux statistiques de l'ensemble des 6-mots.

sont sensiblement identiques. Par conséquent, les Tableaux 6.1, 6.3 et 6.5 donnent une idée des statistiques de quelques 3-mots dans *M1*

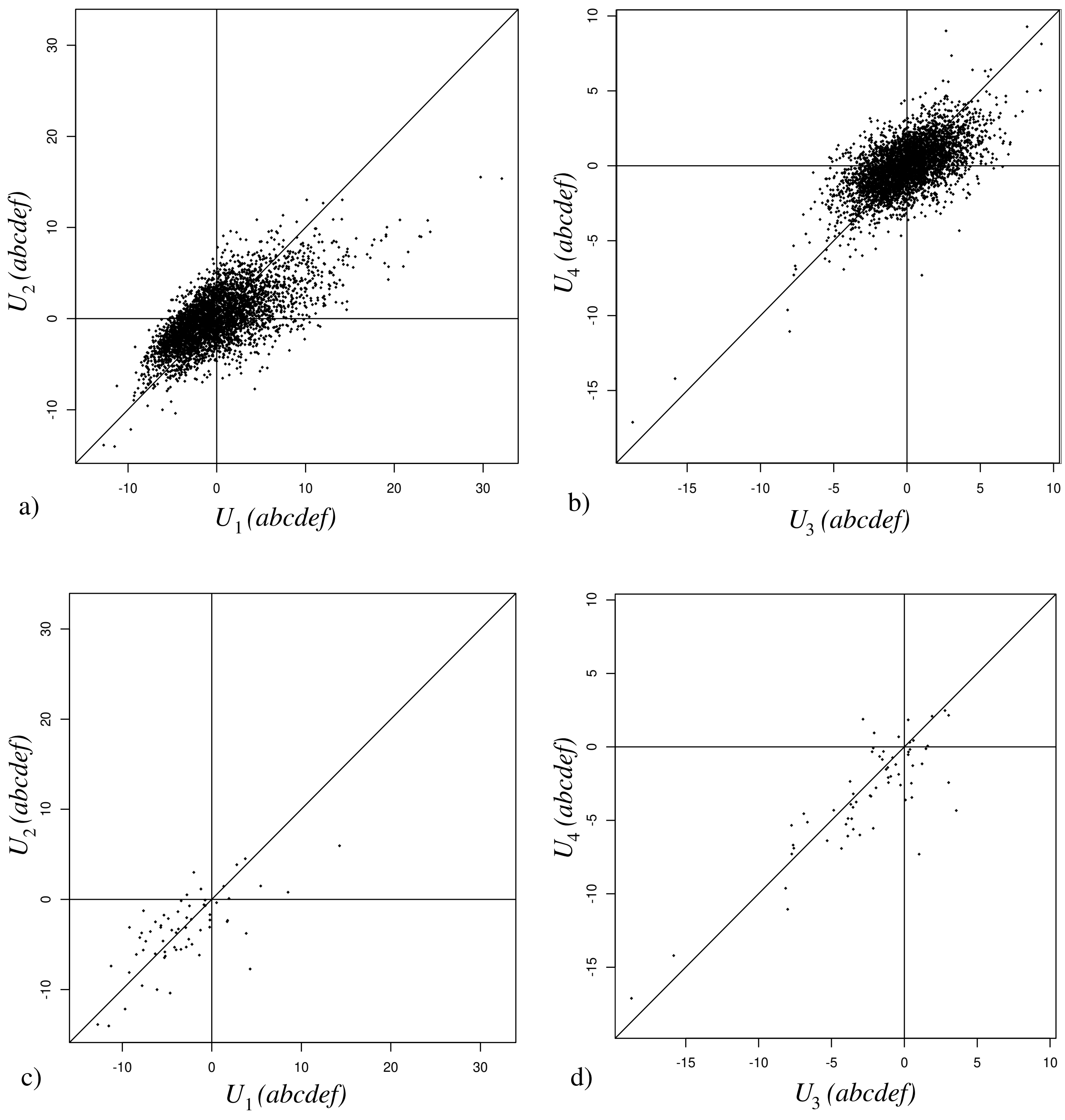


FIG. 8.2 – *Statistiques des 6-mots et des palindromes sous M1 à M4 dans E. coli.*

Les statistiques dans M1 des 6-mots a) et des 6-palindromes c) sont représentées contre celles dans M2 ; les statistiques dans M3 des 6-mots b) et des 6-palindromes d) sont représentées contre celles dans M4.

La plupart des palindromes sont sous-représentés dans les quatre modèles et les mots les plus sous-représentés sont des palindromes. Ceci est d'autant plus visible que l'ordre du modèle est élevé : comme le montre le Tableau 8.1, le nombre de palindromes parmi les 23 mots les plus sous-représentés augmente avec l'ordre du modèle.

TAB. 8.1 – LES 6-MOTS LES PLUS SOUS-REPRÉSENTÉS SOUS CHAQUE MODÈLE DANS *E. coli*.

M1	*M2*	*M3*	*M4*	Rangs
• GGC (-13)	• GCC (-14)	• GGC (-19)	• GGC (-17)	1
• GCC (-12)	• GGC (-14)	• GCC (-16)	• GCC (-14)	2
◦ GCA (-11)	• CGG (-12)	◦ AGC (-8)	⋆ CTG (-11)	3
• CGG (-10)	⋆ CTG (-10)	⋆ CTG (-8)	◦ AGC (-10)	4
⋆ ttggaa (-9)	◦ TGG (-10)	GGT (-8)	TCC (-7)	5
◦ gcttgg (-9)	⋆ CCG (-10)	• CGG (-8)	• CGG (-7)	6
CAC (-9)	tggccg (-9)	◦ CAG (-8)	GAG (-7)	7
TTG (-9)	⋆ ttggaa (-9)	⋆ CCG (-8)	⋆ CCG (-7)	8
cgcatg (-9)	◦ gcttgg (-8)	⋆ ttggaa(-7)	◦ CAG (-7)	9
cttgga (-9)	tttgga (-8)	◦ TGG (-7)	◦ GCA (-6)	10
–	CAC (-8)	CAC (-7)	–	11
–	–	–	CCC (-6)	12
–	–	–	GTC (-6)	13
–	CAG (-8)	–	–	14
GGG (-8)	–	–	CGT (-6)	15
–	–	–	AAA (-6)	16
–	–	–	GGT (-5)	17
–	–	–	–	18
–	–	–	ATC (-5)	19
–	–	–	CAC (-5)	20
–	GCA (-7)	–	–	21
–	–	–	AAG (-5)	22
ATG (-8)	tccaaa (-7)	ttccga (-5)	GGA (-5)	23
8	9	10	19	# de palindromes

Les 6-mots les plus sous-représentés dans E. coli sont listés ici avec, entre parenthèses, leurs statistiques dans chaque modèle. Les 6-palindromes sont notés par leurs trois premières lettres. Les mots non identifiés (–) ne sont pas des palindromes. Les rangs indiqués sont ceux des mots parmi les mots sous-représentés. Dans les dix premières lignes du tableau, certains mots apparaissent exceptionnels sous plus d'un modèle : trois mots sont exceptionnels sous les quatre modèles (marqués avec •), trois *mots sont exceptionnels sous trois modèles (⋆) et cinq mots sont exceptionnels sous deux modèles (∘).*

Le fait que ces palindromes soient sous-représentés dans plus d'un modèle signifie que leur exceptionnelle rareté est obtenue non seulement à partir de sous-mots évités mais aussi en choisissant correctement les lettres qui bordent ces sous-mots.

De plus, le Tableau 8.2 montre que la plupart des 12 6-mots les plus sous-représentés (tous des palindromes sauf CTGGAAG) sont des mots sur-représentés à une lettre près ; seuls les mots de rangs 6, 8 et 10 du Tableau 8.1 ($M4$) n'apparaissent pas.

TAB. 8.2 – LES 6-MOTS SOUS-REPRÉSENTÉES SOUS M4 QUI DIFFÈRENT D'UNE LETTRE DES 6-MOTS SUR-REPRÉSENTÉS DANS *E. coli*.

Les 6-mots sur-représentés [rang de U_4]		Les 6-mots sous-représentés différant d'une lettre [rang de $-U_4$]	
TCCGGC	[1]	TCCGGa [5]	gCCGGC [2]
AGCGCC	[2]	AGCGCt [4]	gGCGCC [1]
GCCGGA	[3]	GCCGGc [2]	tCCGGA [5]
GGCGCT	[4]	GGCGCc [1]	aGCGCT [4]
GAGCTG	[5]	GAGCTc [7]	cAGCTG [9]
TTGCAG	[6]	–	cTGCAG [3]
GCCGGG	[7]	GCCGGc [2]	cCCGGG [12]
CTAGTA	[8]	–	–
TTGAAG	[9]	–	cTGAAG[(a)] [11]
CAGCTT	[10]	CAGCTg [9]	aAGCTT [4]
GGCGCG	[11]	GGCGCc [1]	–

[(a)] *n'est pas un palindrome mais apparaît dans la liste des mots les plus sous-représentés sous $M4$. Les lettres minuscules indiquent les bases mutées.*

Par conséquent, le modèle maximal d'ordre 4 peut être interprété lui-même comme un code complexe sur les 6-mots qui évite une liste de 6-palindromes et de quelques autres mots, et qui au contraire favorise l'occurrence de motifs obtenus en mutant une lettre de ces palindromes évités (Tableau 8.2).

Cette balance entre mots évités et mots préférés qui diffèrent par une seule lettre est partiellement vérifiée dans les modèles d'ordre inférieur comme le montre le Tableau 8.3.

TAB. 8.3 – LES 6-MOTS SOUS- ET SUR-REPRÉSENTÉS QUI DIFFÈRENT D'UNE LETTRE, DANS *E. coli.*

Modèle	Mots sur-représentés [rang de U]	Mots sous-représentés [rang de $-U$]
M1	GCCAGC [3]	GCCgGC [2]
	GCTGGC [7]	GCcGGC [2]
M2	TCGCCA [4]	TgGCCA [5]
	TGGCGA [7]	TGGCcA [5]
M3	GCCGGA [1]	GCCGGc [2]
	TCCGGC [4]	gCCGGC [2]
	TCGCCA [7]	TgGCCA [10]
	AGGCCG [10]	cGGCCG [6]

Les lettres minuscules indiquent les lettres mutées.

• **Étude des 8-palindromes sous** *M1* **à** *M6*

Si l'on considère la distribution des statistiques des 8-palindromes parmi l'ensemble des statistiques des 8-mots, on remarque que beaucoup de 8-palindromes sont évités sous *M1* et *M2*. Ceci reste cependant moins vrai lorsque l'ordre du modèle augmente. Sous *M5* et *M6*, les statistiques des palindromes sont même dispersées dans l'éventail des statistiques des 8-mots (non montré). Dans le Tableau 8.4, le rang (entre crochets) des palindromes sous-représentés montre que les palindromes ne sont pas majoritairement parmi les mots les plus évités.
Finalement, la plupart de ces palindromes sont des extensions de 6- ou 4-palindromes sous-représentés dans le même modèle (entre parenthèses dans le Tableau 8.4). Seuls deux 8-palindromes sont évités après que six ou sept de leurs lettres aient été choisies. Cela signifie que les deux palindromes CAGATCTG et CAGGCCTG peuvent avoir une fonction biologique caractérisée par le mot dans son ensemble et pas seulement par une partie du mot. Pour les autres palindromes, au contraire, les sous-représentations de CGGCGCCG et TGGCGCCA, par exemple, correspondent au fait que GGCGCC est évité et que de plus, la présence d'un C et d'un G ou, d'un T et d'un A, à ses extrémités est évitée.

TAB. 8.4 – LES 8-PALINDROMES SOUS-REPRÉSENTÉS DANS *E. coli*.

M1	*M2*	*M3*	*M4*	*M5*	*M6*
C(GGC) [3]	C(GGC) [2]	C(GCC) [3]	T(GGC) [3]	CAGA [21]	CAGG [24]
T(GCA) [6]	C(GCC) [8]	T(GGC) [5]	C(GCC) [6]	CAGG [36]	
C(GCC) [14]	T(GGC) [13]	C(GGC) [7]	C(GGC) [10]		
C(GCA) [24]	CA(GA) [32]	T(GCC) [41]	T(GCC) [16]		
G(CAC) [28]	G(CGG) [35]	C(AGC) [71]	CA(GG) [18]		
TT(GG) [78]	G(CCG) [57]	A(GGC) [73]	C(AGC) [62]		
T(GGC) [89]	CT(GA) [75]	C(TGG) [105]			
	G(GCC) [91]				
	T(CAG) [99]				
	G(CTG) [101]				

Pour chacun des modèles M1 à M6, on mentionne ici les 8-palindromes ayant une statistique $U_m < -3$, avec leur rang entre crochets, sachant que le mot le plus sous-représenté est de rang 1. Ils sont représentés par leurs quatre premières lettres. Les 6-palindromes sous-représentés du Tableau 8.1 et les 4-palindromes sous-représentés, qui apparaissent comme sous-mots sont mis entre parenthèses. Deux 8-palindromes soulignés sont une double répétition d'un 4-palindrome sous-représenté.

La caractérisation statistique de la sous-représentation des 8-palindromes est par conséquent différente de celles des 6-palindromes, et montre à nouveau une fonction spécifique des 6-palindromes dans *E. coli*.

8.2 Les 4-, 6-, et 8-palindromes dans les phages *Lambda* et *T7*

Étude des palindromes dans le phage *Lambda*

L'étude de la fréquence des palindromes dans le génome du phage *Lambda* (48502 bp) donne des résultats assez semblables à ceux obtenus chez *E. coli* tout en étant un peu moins spectaculaires.

- la plupart des 4-palindromes sont sous-représentés dans au moins un des deux modèles *M1* et *M2*, et seul CCGG est sur-représenté (Figure 8.3).

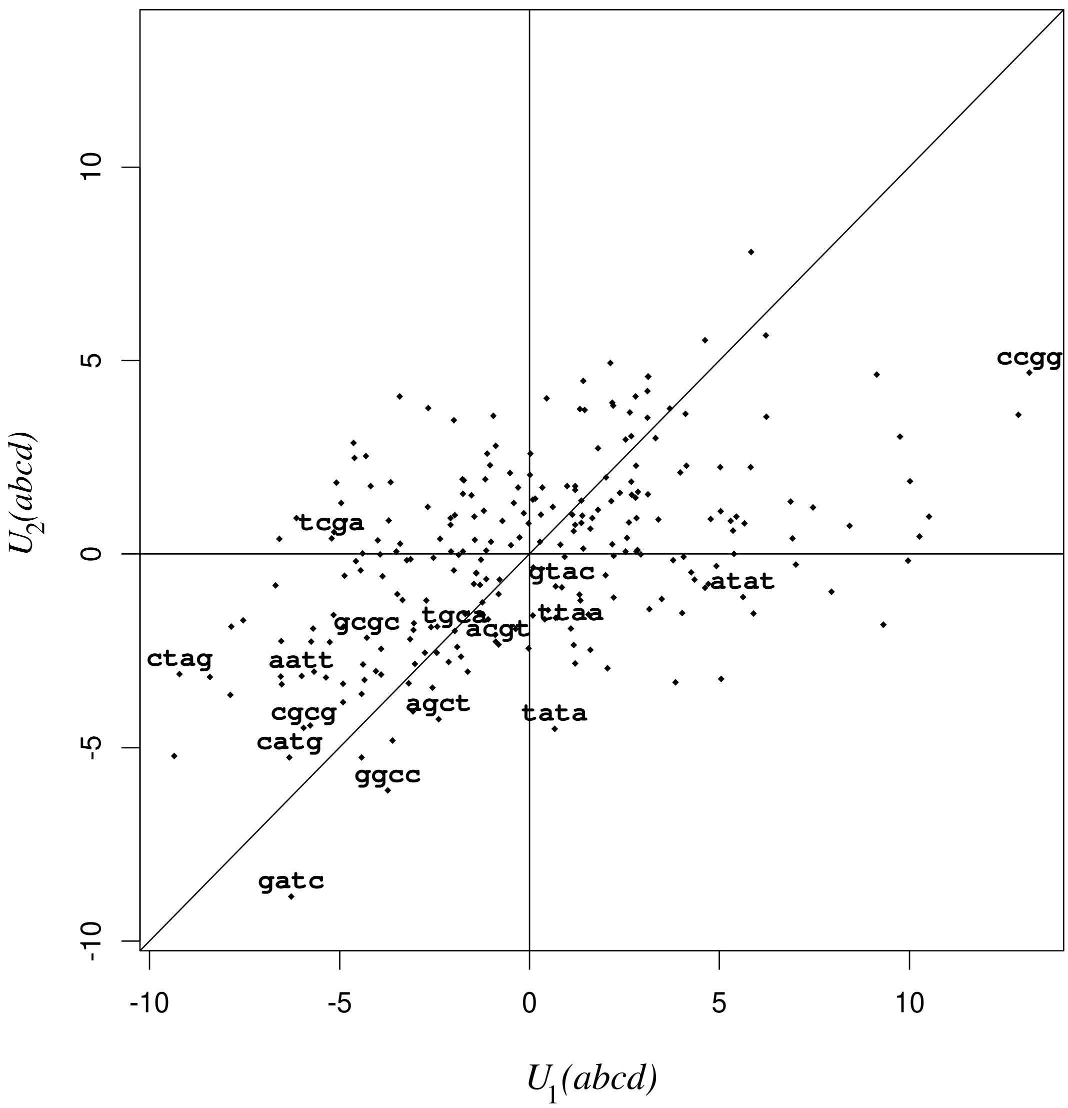


FIG. 8.3 – *Statistiques des 4-mots sous M1 et M2 dans le phage Lambda.*

Si on compare les Figures 8.1 et 8.3, la position des palindromes dans les deux nuages de points sont sensiblement les mêmes, à l'exception toutefois de CCGG et GATC qui se démarquent nettement du nuage pour le phage *Lambda*.

- Dans l'ensemble, les palindromes de longueur 6 sont aussi sous-représentés dans les quatre modèles dans le phage *Lambda* (Figure 8.4).

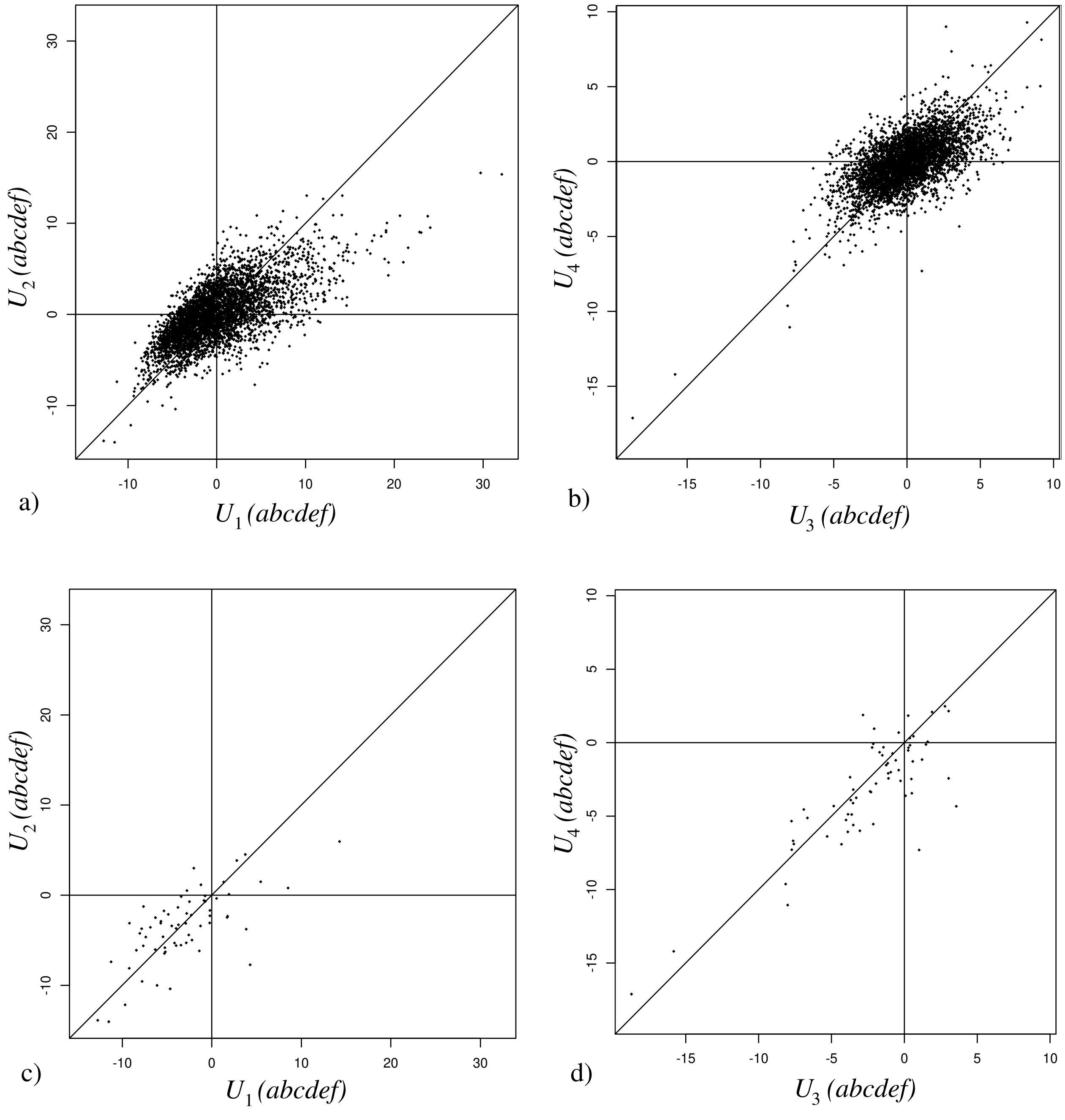


FIG. 8.4 – *Statistiques des 6-mots et des palindromes sous M1 à M4 dans le phage Lambda.*

Les statistiques dans M1 des 6-mots a) et des 6-palindromes c) sont représentées contre celles dans M2 ; les statistiques dans M3 des 6-mots b) et des 6-palindromes d) sont représentées contre celles dans M4.

Cependant, les statistiques des palindromes, et même de l'ensemble des 6-mots, sont moins élevées que celles obtenues chez *E. coli*; ceci est vraisemblablement dû à une plus petite longueur de séquence pour le phage *Lambda*. Certes, les 6-mots les plus exceptionnellement rares sont souvent des palindromes, mais la présence des 6-palindromes parmi les mots les plus sous-représentés dans les modèles *M1*, *M3* et *M4* est nettement moins importante que celle observée chez *E. coli*. Le Tableau 8.5 donne le rang des 6-palindromes possédant une statistique inférieure à -3, pour chacun des quatre modèles.

TAB. 8.5 – LES 6-PALINDROMES LES PLUS SOUS-REPRÉSENTÉS SOUS CHAQUE MODÈLE DANS LE PHAGE *Lambda*.

M1	*M2*	*M3*	*M4*
CGC [2]	<u>GCC</u> [1]	<u>GCC</u> [1]	<u>GCC</u> [1]
<u>GCA</u> [4]	<u>CGG</u> [2]	<u>GGC</u> [2]	<u>GGC</u> [2]
<u>GGC</u> [8]	<u>CAG</u> [4]	<u>CGG</u> [4]	<u>AGC</u> [3]
<u>GCC</u> [9]	<u>GGC</u> [5]	<u>AGC</u> [6]	<u>CCG</u> [6]
GCG [25]	<u>GCA</u> [6]		CCA [7]
CAA [28]	<u>CCG</u> [7]		
ACA [29]	CGC [8]		
AGC [33]	TGA [11]		
<u>CGG</u> [37]	AGC [17]		
<u>GGG</u> [46]			

Les palindromes, notés par leur trois premières lettres, les plus sous-représentés dans le phage Lambda sont listés ici avec, entre crochets, leurs rangs dans chaque modèle. Les palindromes soulignés sont exceptionnellement rares chez E. coli.

Un certain nombre de palindromes (soulignés) du Tableau 8.5 sont aussi évités chez *E. coli* (Tableau 8.1). Cependant, il est intéressant de remarquer que quelques 6-palindromes exceptionnellement rares chez *E. coli* ne sont pas particulièrement sous-représentés chez *Lambda*. C'est le cas de CTGCAG et TGGCCA qui possèdent des statistiques inférieures à 1 en valeur absolue dans les modèles *M2* et *M3*, puis de GAGCTC et CTGCAG dont les statistiques dans le modèles *M4* sont de l'ordre de -1.5.

- De façon globale, les 8-palindromes n'ont pas tendance à être plutôt sous-représentés ou plutôt sur-représentés. Quelques rares 8-palindromes ont une statistique de l'ordre de 3 dans *M1*, *M5* ou *M6*, mais aucun des 8-mots ne possède de statistiques inférieures à -3 dans chacun des modèles *M1* à *M6*.

Étude des palindromes dans le phage *T7*

Dans l'ensemble, les palindromes de longueur 4 et 6 sont sous-représentés dans le génome de *T7*. Toutefois, on peut noter quelques différences avec *E. coli* et *Lambda*.

- Parmi les 4-palindromes, seul TTAA est sur-représenté, dans le modèle *M1*.
- GATC est le 4-mot le plus sous-représenté dans les modèles *M1* et *M2* et se démarque énormément du nuage composé des autres 4-mots.
- CCGG n'est pas sur-représenté dans *T7*, contrairement à *E. coli* et *Lambda*.
- Les 6-palindromes sont majoritairement sous-représentés, mais les valeurs prises par les statistiques des 6-mots sont plus faibles, en valeur absolue, que celles obtenues pour *E. coli* et *Lambda*. En particulier, le nuage représentant le modèle *M3* versus *M4* ne possèdent pas de "traîne" vers les valeurs de statistiques très négatives, comme on a pu le remarquer sur les Figures 8.2 et 8.4.
- Le nombre de 6-palindromes sous-représentés communs à *T7* et *E. coli* est nettement inférieur à celui obtenu en comparant *E. coli* et *Lambda* (Tableau 8.5).
- Les trois palindromes de longueur 6 particulièrement évités chez *E. coli* (sous-représentés dans les quatre modèles, Tableau 8.1), ne le sont pas chez *T7*.

Chapitre 9

Étude du motif GCTGGTGG dans *E. coli* et les phages *Lambda* et *T7*

9.1 Fréquence de GCTGGTGG dans *E. coli*

L'étude ci-dessous a été réalisée sur la même séquence d'ADN de *E. coli* utilisée au Chapitre 8 à savoir, les 384243 dernières bases, situées à droite de l'origine de réplication, du contigue formé des quatre séquences L10328, M87049, L19201 et U00006 (Genbank). Elle montre que la fréquence du motif GCTGGTGG, appelé Chi d'*E. coli* (Crossover hotspot instigator) présente des propriétés statistiques particulières.

Auparavant, nous décrivons brièvement le rôle biologique de Chi. Pour contrôler l'entrée d'ADN exogène, *Escherichia coli* dispose, notamment, d'un système utilisant une enzyme multifonctionnelle appelée RecBCD. Cette enzyme, principale exonucléase d'*E. coli*, est activée par l'intermédiaire d'une courte séquence d'ADN, appelée Chi, qui est répartie de façon fréquente sur le chromosome d'*E. coli*. Plusieurs modèles d'interactions entre Chi et RecBCD ont été proposés et leur principe est le suivant (Figure 9.1). RecBCD se fixe à une extrémité libre de l'ADN double brin (9.1-A) et progresse en désappariant les deux brins. L'ADN est ensuite dégradé de façon asymétrique par l'activité exonucléasique de RecBCD, jusqu'à ce que cette dernière rencontre un site Chi correctement orienté (9.1-B). La reconnaissance de ce site inactive la fonction exonucléasique de RecBCD (9.1-C). Une coupure est ensuite introduite sur le brin portant Chi, quelques nucléotides avant le site (du côté 3'), puis l'enzyme poursuit sa progression en séparant les deux brins de la molécule d'ADN avec son activité hélicase (9.1-D). Ceci produit ainsi une extrémité 3' recombinogène.
La présence de nombreux sites Chi sur le chromosome est donc une protection de l'ADN d'*E. coli* contre la dégradation par l'enzyme RecBCD, et un stimulant à la recombinaison provoquée par la rencontre entre Chi et RecBCD.

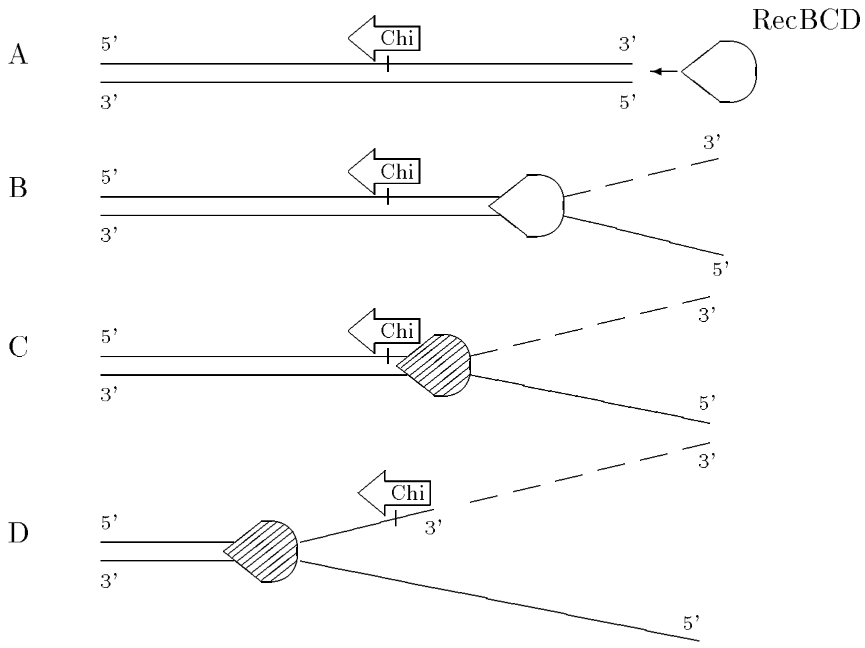


FIG. 9.1 – *Modèle d'interaction entre RecBCD et Chi.*

Afin d'étudier la fréquence du motif Chi=GCTGGTGG dans la séquence d'*E. coli*, nous avons ajusté successivement chacun des modèles *M0* , *M1* , ... et *M6*, et calculé les statistiques U asymptotiquement gaussiennes centrées réduites correspondantes (Partie I) ; le modèle *M0* est celui où les bases de la séquence sont supposées indépendantes.

Le Tableau 9.1 montre que Chi est le 8-mot le plus sur-représenté lorsque l'on ajuste l'un des trois modèles *M0*, *M1* et *M2* sur la séquence. De plus, il reste parmi les 8-mots les plus sur-représentés lorsque l'on augmente l'ordre du modèle markovien. Le fait que Chi soit exceptionnellement fréquent dans chaque modèle traduit donc une forte contrainte vis à vis de tous ses sous-mots de longueur 2 à 7 car le nombre de GCTGGTGG est toujours plus important que celui prédit par les di-, tri-, tétra-nucléotides, etc ...

TAB. 9.1 – STATISTIQUES DE GCTGGTGG SOUS LES MODÈLES M0 À M6 DANS *E. coli.*

W =GCTGGTGG, $N(W) = 81$				
Modèle	$\widehat{\mathbb{E}}N(W)$	$\widehat{\mathrm{Var}}(N(W))$	U(W)	Rang
M0	7.717	7.708	26.3963	1
M1	7.690	7.673	26.4654	1
M2	19.288	18.994	14.1600	1
M3	32.945	31.318	8.5870	28
M4	32.412	28.548	9.0938	18
M5	38.614	26.959	8.1634	7
M6	62.404	18.978	4.2688	34

Le rang est celui de Chi parmi les mots sur-représentés.

Lorsque l'on s'intéresse au motif CCACCAGC, qui est le conjugué[1] de Chi, on remarque (Tableau 9.2) qu'il ne fait pas partie des 8-mots les plus sur-représentés pour chacun des sept modèles ajustés sur la séquence. Cela signifie que la sur-représentation du motif Chi n'est effective que sur un brin d'ADN et non sur les deux brins complémentaires (dissymétrie).

TAB. 9.2 – STATISTIQUES DE CCACCAGC SOUS LES MODÈLES M0 À M6 DANS *E. coli.*

W =CCACCAGC, $N(W) = 17$				
Modèle	$\widehat{\mathbb{E}}N(W)$	$\widehat{\mathrm{Var}}(N(W))$	$U(W)$	Rang
M0	5.775	5.769	4.6734	2399
M1	4.768	4.760	5.6067	913
M2	12.132	11.989	1.4060	10534
M3	18.187	17.507	-0.2836	34055
M4	17.051	15.439	-0.0129	29231
M5	18.477	13.885	-0.3964	38648
M6	21.813	8.648	-1.6367	61431

Le rang 1 correspond à celui du 8-mot le plus sur-représentés.

Pour apprécier la sur-représentation de Chi, nous avons effectué la même analyse pour trois motifs de fréquence aussi élevée dans la séquence que celle de Chi (81 occurrences). Ces trois motifs, GGCGCTGG, CGCTGGCG et GCCAGCAG sont en fait, avec Chi, les mots les plus sur-représentés sous les modèles *M0* ou *M1*. Cependant, le Tableau 9.3 montre que ces motifs

1. défini au Chapitre 8

ne possèdent pas les propriétés statistiques de Chi que nous avons décrites ci-dessus. En effet, chacun de ces trois mots perdent leur caractère exceptionnel dans les modèles d'ordre supérieur ou égal à 3 (sauf GGCGCTGG qui est 49ème dans *M4*). De plus, ils ne présentent pas de dissymétrie quant au brin d'ADN sur lequel on l'étudie, puisque leurs conjugués possèdent des statistiques du même ordre.

TAB. 9.3 – STATISTIQUES DE GGCGCTGG, CGCTGGCG, GCCAGCAG ET DE LEURS CONJUGUÉS, SOUS LES MODÈLES M0 À M6, DANS *E. coli*.

	W =GGCGCTGG $N(W)=77$		W =CGCTGGCG $N(W)=68$		W =GCCAGCA $N(W)=57$	
Modèle	$U(W)$	Rang	$U(W)$	Rang	$U(W)$	Rang
M0	24.5041	**2**	22.2126	**3**	19.6367	8
M1	19.2064	**2**	14.2766	13	18.5421	**3**
M2	11.2313	6	7.4762	78	8.2979	49
M3	6.0286	112	1.7250	6896	5.2716	208
M4	7.4302	49	0.2647	23754	2.0605	4043
M5	3.4283	577	0.6474	17247	3.5052	521
M6	-0.4911	42656	0.0313	30053	1.7435	4558
	$\overline{W}$ =CCAGCGCC $N(\overline{W})$ =57		$\overline{W}$ =CGCCAGCG $N(\overline{W})=57$		$\overline{W}$ =CTGCTGGC $N(\overline{W})=43$	
Modèle	$U(\overline{W})$	Rang	$U(\overline{W})$	Rang	$U(\overline{W})$	Rang
M0	19.9410	7	19.1224	11	13.9432	43
M1	16.8332	6	13.2105	25	11.1468	58
M2	9.7047	17	6.7799	125	3.9364	1319
M3	5.6923	152	1.9433	5625	1.2880	10354
M4	7.8785	37	0.0855	27192	-0.6570	44057
M5	3.6847	429	-0.6240	44029	0.8853	13609
M6	1.9955	3062	-0.6590	46232	0.4417	21654

Les mots GGCGCTGG, CGCTGGCG et GCCAGCAG (avec Chi) sont les 8-mots les plus sur-représentés sous les modèles M0 ou M1. Le rang 1 correspond à celui du 8-mot le plus sur-représenté.

9.2 Fréquence de GCTGGTGG dans les gènes de *E. coli*

Pour étudier la fréquence de Chi dans les gènes de *E. coli*, nous avons extrait de la séquence de 384243 bp uniquement les gènes qui codent sur ce brin. Nous avons ainsi obtenu une séquence codante de 182865 bp, dans laquelle 66 des 81 occurrences du motif Chi sont encore présentes. Nous avons ajusté sur cette séquence codante des modèles markoviens dont les transitions sont périodiques de période 3, pour prendre en compte la position des nucléotides par rapport aux codons, que l'on appelle la *phase*. Le Tableau 9.4 nous montre que Chi reste exceptionnellement fréquent dans les modèles *M0_3* à *M5_3* mais n'est plus exceptionnel dans *M6_3*.

De plus, le caractère exceptionnel de sa fréquence se manifeste essentiellement sur la phase 1, c'est à dire lorsqu'il est positionné G|CTG|GTG|G par rapport aux codons. Toutefois, dans le modèle *M1_3*, il apparaît également sur-représenté sur la phase 3, c'est à dire dans la position GC|TGG|TGG. Il faut dire que le codon TGG est exceptionnellement fréquent dans ce modèle, ce qui semble en être une explication puisque dans le modèle *M2_3*, qui s'ajuste sur la fréquence des codons, Chi n'est plus exceptionnel en phase 3.

Le motif GCTGGTGG est sur-représenté dans la séquence codante de *E. coli* étudiée, lorsque l'on ajuste un modèle de chaîne de Markov à transitions 3-périodiques. Cette sur-représentation, notamment dans le modèle *M2_3*, est ainsi plus forte que la contrainte liée au codage en protéines et qui induit, on le sait, une préférence de codons. La phase 1 semble toutefois être préférée aux deux autres.

TAB. 9.4 – STATISTIQUES SUR CHACUNE DES TROIS PHASES ET GLOBALE, DE GCTGGTGG SOUS LES MODÈLES M0_3 À M6_3.

Modèle		Comptage	Espérance	Variance	Statistique	Rang
	globale	66	6.671	6.651	23.0049	**2**
M0_3	phase 1	58	5.195	5.176	23.2104	**2**
	phase 2	3	1.251	1.249	1.5647	7652
	phase 3	5	0.226	0.226	10.0480	128
	globale	66	9.635	9.562	18.2281	**1**
M1_3	phase 1	58	7.530	7.460	18.4782	**2**
	phase 2	3	1.993	1.986	0.7147	13083
	phase 3	5	0.113	0.113	14.5549	**10**
	globale	66	22.735	21.853	9.2551	**1**
M2_3	phase 1	58	19.042	18.207	9.1301	**12**
	phase 2	3	2.167	2.143	0.5687	13297
	phase 3	5	1.525	1.504	2.8339	1855
	globale	66	33.801	30.442	5.8359	**9**
M3_3	phase 1	58	28.719	25.609	5.7863	108
	phase 2	3	2.720	2.633	0.1726	17001
	phase 3	5	2.362	2.200	1.7783	4421
	globale	66	36.505	28.471	5.5278	**22**
M4_3	phase 1	58	32.288	24.696	5.1740	124
	phase 2	3	1.827	1.718	0.8945	9647
	phase 3	5	2.389	2.057	1.8204	3848
	globale	66	40.378	23.120	5.3288	**12**
M5_3	phase 1	58	35.712	19.737	5.0169	52
	phase 2	3	1.280	1.119	1.6262	4274
	phase 3	5	3.386	2.264	1.0728	7799
	globale	66	55.161	13.991	2.8978	231
M6_3	phase 1	58	48.763	11.754	2.6943	240
	phase 2	3	2.065	1.136	0.8775	8724
	phase 3	5	4.333	1.101	0.6355	11238

Dans chaque modèle Mm_3, on donne tout d'abord le nombre d'occurrences de Chi, une estimation de son espérance et de sa variance, ainsi que la statistique asymptotiquement gaussienne centrée réduite associée (globale), puis on donne la valeur de ces quantités pour chacune des trois phases. Le rang est celui parmi les mots sur-représentés.

9.3 Fréquence de GCTGGTGG dans les phages *Lambda* et *T7*

- Le motif GCTGGTGG est absent dans le génome du phage *Lambda* et sa fréquence n'est pas statistiquement exceptionnelle. Aucun des mots obtenus en mutant une de ses lettres n'est exceptionnellement fréquent. De plus, l'étude de la fréquence de son conjugué CCACCAGC, qui revient à étudier Chi sur le brin complémentaire, aboutit aux mêmes conclusions. Par ailleurs, parmi les 8-mots exceptionnellement fréquents dans *Lambda*, aucun ne l'est dans les six modèles *M1* à *M6* à la fois.

- Le génome du phage *T7* comporte six occurrences de GCTGGTGG. Ce motif est sur-représenté dans le modèle *M1* mais cela semble être expliqué par le fait que ses deux sous-mots de longueur sept GCTGGTG et CTGGTGG sont exceptionnellement fréquents sous ce modèle. Dans les modèles d'ordre supérieur, le motif GCTGGTGG n'est plus de fréquence exceptionnelle. Certains des mots obtenus à partir d'une mutation d'une de ses lettres sont sur-représentés mais ne font pas partie des 8-mots les plus exceptionnellement fréquents. Son conjugué CCACCAGC n'est pas présent le long du génome du phage *T7*. Toutefois, lorsque l'on compare les 8-mots les plus exceptionnellement fréquents dans les six modèles, on s'aperçoit que deux d'entre eux sont communs aux six modèles : CGACTCAC et TAATACGA (respectivement 12 et 14 occurrences). De plus, leurs conjugués sont absents dans la séquence et possèdent des statistiques non significatives.

Il apparaît ainsi clairement que le motif GCTGGTGG ne possède pas les mêmes propriétés statistiques, quant à sa fréquence, lorsqu'on l'étudie dans le génome des phages *Lambda* et *T7*, et dans *E. coli*. GCTGGTGG semble être ainsi spécifique à la bactérie *E. coli*. Cependant, les deux motifs CGACTCAC et TAATACGA possèdent sur le génome de *T7* les mêmes propriétés statistiques que le Chi de *E. coli*, à savoir une fréquence exceptionnellement élevée, dans tous les modèles, et uniquement sur un brin d'ADN. On peut donc se poser la question "Sont-ils les Chi de *T7*?" mais seule la Biologie peut apporter une réponse.

Quatrième partie

Appendices

Annexe A

Compléments au Chapitre 2 : les modèles M2 et M0

A.1 Asymptotique des rapports R_k de cardinaux

Preuve du Lemme 1 : Pour établir le comportement asymptotique de R_k, nous étudions en détail le cas où $k = 1$, puis nous adapterons la méthode pour k quelconque.

On utilise la formule de Whittle (Théorème 2)

$$\frac{\#(\mathbb{S}\backslash\mathbb{W})}{\#\mathbb{S}} = \frac{K(\mathbb{S}\backslash\mathbb{W})}{K(\mathbb{S})} \frac{H_{X'_{n-1}}(\mathbb{S}\backslash\mathbb{W})}{H_{X'_{n-1}}(\mathbb{S})},$$

puis on fait un développement asymptotique de chacun des deux quotients.

• Développement asymptotique de $K(\mathbb{S}\backslash\mathbb{W})/K(\mathbb{S})$.

$$\begin{aligned}\frac{K(\mathbb{S}\backslash\mathbb{W})}{K(\mathbb{S})} &= \prod_A \frac{N_{A+}(\mathbb{S}\backslash\mathbb{W})!}{N_{A+}!} \prod_{A,B} \frac{N_{AB}!}{N_{AB}(\mathbb{S}\backslash\mathbb{W})!} \\ &= \prod_A \frac{N_{A+}(\mathbb{S}\backslash\mathbb{W})!}{N_{A+}!} \prod_{\substack{AB \\ \text{compatibles}}} \frac{N_{AB}!}{N_{AB}(\mathbb{S}\backslash\mathbb{W})!} \cdot \left(\frac{N_{UV}!}{N_{UV}(\mathbb{S}\backslash\mathbb{W})!}\right)^{1-\delta}. \qquad (A.1)\end{aligned}$$

La dernière égalité vient du fait que si AB est différent de UV et n'est pas compatible, alors $N_{AB} = N_{AB}(\mathbb{S}\backslash\mathbb{W}) = 0$. De plus, lorsque UV n'est pas compatible ($\delta = 0$), on a $N_{UV} = 0$ et $N_{UV}(\mathbb{S}\backslash\mathbb{W}) = 1$; c'est pourquoi, comme $N_{A+}(\mathbb{S}\backslash\mathbb{W}) = N_{A+} - d_{A+}$ et $N_{AB}(\mathbb{S}\backslash\mathbb{W}) = N_{AB} - d_{AB}$, on obtient

$$\frac{K(\mathbb{S}\backslash\mathbb{W})}{K(\mathbb{S})} = \prod_A \frac{(N_{A+} - d_{A+})!}{N_{A+}!} \prod_{\substack{AB \\ \text{compatibles}}} \frac{N_{AB}!}{(N_{AB} - d_{AB})!}. \qquad (A.2)$$

Si $D \geq 0$ et N tend vers l'infini, on a le développement asymptotique suivant

$$\frac{N!}{(N-D)!} = N^D \left(1 - \frac{D(D-1)}{2N} + \mathrm{o}\left(\frac{1}{N}\right)\right) ;$$

or, N_{AB} tend vers l'infini dès que AB est compatible, $d_{A+} \geq 0$ pour tout A, et d_{AB} est strictement négatif uniquement si $AB = UV$ et $n_{UV} = 0$. Par conséquent, si $n_{UV} > 0$, on a

$$\begin{aligned}
\frac{K(\mathbb{S}\backslash\mathbb{W})}{K(\mathbb{S})} &= \prod_A \frac{\prod_{B|AB\ comp}(N_{AB})^{d_{AB}}}{(N_{A+})^{d_{A+}}} \\
&\times \left(1 - \sum_{AB comp} \frac{d_{AB}(d_{AB}-1)}{2N_{AB}} + \sum_A \frac{d_{A+}(d_{A+}-1)}{2N_{A+}} + \mathrm{o}\left(\frac{1}{n}\right)\right) . \qquad (A.3)
\end{aligned}$$

En fait, si $n_{UV} = 0$ alors que UV est compatible (aucun mot non compatible n'est pris en compte dans (A.2)), on obtient le même développement asymptotique que (A.3) en employant une démarche similaire, avec toutefois $N_{UV}!/(N_{UV} - d_{UV})! = N_{UV}^{-1}$ car $d_{UV} = -1$.
Or,

$$\begin{array}{lll}
d_{A+} = \sum_{B|ABcomp} d_{AB}, & \forall A\,, & \text{si } UV \text{ compatible,} \\
d_{A+} = \sum_{B|ABcomp} d_{AB}, & \forall A \neq U\,, & \text{si } UV \text{ n'est pas compatible,} \\
d_{U+} = \sum_{B|UBcomp} d_{UB} - 1\,, & & \text{si } UV \text{ n'est pas compatible,}
\end{array}$$

et le premier facteur de (A.3) devient

$$\prod_A \frac{\prod_{B|AB\ comp}(N_{AB})^{d_{AB}}}{(N_{A+})^{d_{A+}}} = (N_{U+})^{1-\delta} \prod_{ABcomp} \left(\frac{N_{AB}}{N_{A+}}\right)^{d_{AB}} = (N_{U+})^{1-\delta}\, C\,.$$

Ainsi, en remplaçant dans (A.3), on obtient

$$\frac{K(\mathbb{S}\backslash\mathbb{W})}{K(\mathbb{S})} = (N_{U+})^{1-\delta}\, C \left(1 - \sum_{ABcomp} \frac{d_{AB}(d_{AB}-1)}{2N_{AB}} + \sum_A \frac{d_{A+}(d_{A+}-1)}{2N_{A+}} + \mathrm{o}\left(\frac{1}{n}\right)\right) . \qquad (A.4)$$

Lorsque l'on supprime l'intérieur de k mots $\mathbb{W}$ pour R_k, chaque d_{AB} est remplacé par kd_{AB}, les facteurs C et N_{U+} sont changés en leur puissance k-ième et le facteur $(1/k!)^{1-\delta}$ apparaît du fait que dans la formule correspondante à (A.1), $N_{UV}(\mathbb{S}\backslash\mathbb{W}\cdots\backslash\mathbb{W}) = k$.

- En considérant maintenant le rapport $H_{X'_{n-1}}(\mathbb{S}\backslash\mathbb{W})\ /\ H_{X'_{n-1}}(\mathbb{S})$, on remarque que

$$\begin{cases}
\forall AB \text{ compatible,} & \left(\widehat{\Pi}(A,B) \mid \mathbb{S}\backslash\mathbb{W}\right) = \widehat{\Pi}(A,B)\left(1 + \dfrac{D_{AB}}{n} + \mathrm{o}(\dfrac{1}{n})\right) \\
& \text{avec } D_{AB} = \dfrac{n}{N_{A+}} d_{A+} + \dfrac{n}{N_{AB}} d_{AB}\,, \\
\forall AB \text{ non compatible, } AB \neq UV\,, & \left(\widehat{\Pi}(A,B) \mid \mathbb{S}\backslash\mathbb{W}\right) = \widehat{\Pi}(A,B) = 0\,, \\
\text{Si } UV \text{ n'est pas compatible,} & \widehat{\Pi}(U,V) = 0 \text{ et } \left(\widehat{\Pi}(U,V) \mid \mathbb{S}\backslash\mathbb{W}\right) = \dfrac{1}{N_{U+}(\mathbb{S}\backslash\mathbb{W})}\,.
\end{cases}$$

Puisque $H_{X'_{n-1}}(\mathbb{S})$ (resp. $H_{X'_{n-1}}(\mathbb{S}\backslash\mathbb{W})$) est une forme polynômiale en les paramètres $\widehat{\Pi}(A,B)$ (resp. $(\widehat{\Pi}(A,B) \mid \mathbb{S}\backslash\mathbb{W})$), si UV est compatible, on a le développement

$$\frac{H_{X'_{n-1}}(\mathbb{S}\backslash\mathbb{W})}{H_{X'_{n-1}}(\mathbb{S})} = 1 + \frac{\Delta_1^{**}}{n} + \mathrm{o}\left(\frac{1}{n}\right)\,, \qquad (A.5)$$

où Δ_1^{**} est linéaire en les variables d_{AB}, et lorsque l'on s'intéresse à R_k, on change d_{AB} en kd_{AB} dans l'expression de D_{AB}, ce qui assure que $\Delta_k^{**} = k\Delta_1^{**}$ par linéarité.
Lorsque UV n'est pas compatible, on se ramène à la situation précédente en prenant soin de choisir $H_{X'_{n-1}}(\mathbb{S})$ et $H_{X'_{n-1}}(\mathbb{S}\backslash\mathbb{W})$, non pas comme étant les cofacteurs de n'importe quel élément de la ligne "X'_{n-1}", mais ceux de l'élément (X'_{n-1}, V) ; de cette façon, on travaille avec des couples (A, B) vérifiant tous

$$\left(\widehat{\Pi}(A,B) \mid \mathbb{S}\backslash\mathbb{W}\right) = \widehat{\Pi}(A,B)\left(1 + \frac{D_{AB}}{n} + \mathrm{o}\left(\frac{1}{n}\right)\right) .$$

En combinant les formules (A.4) et (A.5), on démontre le lemme. □

A.2 Variance conditionnelle asymptotique pour les 4-mots dans M2

On vérifie ici que la limite presque sûre de $\frac{1}{n}\mathrm{Var}(N_{w_1w_2w_3w_4} \mid \mathcal{S}_2)$ est égale à $\sigma^2_{w_1w_2w_3w_4}$, dont l'expression est donnée en (2.8). On va distinguer d'une part, les mots pour lesquels $w_2 = w_3$ ou non, et d'autre part, les mots qui peuvent se chevaucher sur plus de deux lettres (seuls les mots du type $aaaa$ sont dans ce cas). On étudie donc, par exemple, les trois types de mots suivants : $abcd$, $abbc$ et $aaaa$ où les lettres a, b, c, d sont distinctes deux à deux.

• Cas d'un mot du type $abcd$: d'après le Lemme 3, on a

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(N_{abcd} \mid \mathcal{S}_2) = \mu(abcd) + \frac{\mu(abcd)^2}{\mu(ab)}(1-2d_{ab+}) + \mu(abcd)^2\left(\sum_{x,y}\frac{d^2_{xy+}}{\mu(xy)} - \sum_{x,y,z}\frac{d^2_{xyz}}{\mu(xyz)}\right).$$

Or, seuls d_{abc} et d_{bcd} sont non nuls ($d_{abc} = d_{bcd} = 1$) ; par conséquent, seuls d_{ab+} et d_{bc+} sont non nuls et valent 1. Ainsi,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(N_{abcd} \mid \mathcal{S}_2) &= \mu(abcd) - \frac{\mu(abcd)^2}{\mu(ab)} + \mu(abcd)^2\left(\frac{1}{\mu(ab)} + \frac{1}{\mu(bc)} - \frac{1}{\mu(abc)} - \frac{1}{\mu(bcd)}\right) \\
&= \frac{\mu(abcd)}{\mu(bc)}\left(\mu(bc) + \mu(abcd) - \frac{\mu(abcd)\mu(bc)}{\mu(abc)} - \frac{\mu(abcd)\mu(bc)}{\mu(bcd)}\right) \\
&= \frac{\mu(abcd)}{\mu(bc)}\left(\mu(bc) + \frac{\mu(abc)\mu(bcd)}{\mu(bc)} - \mu(bcd) - \mu(abc)\right) \\
&= \frac{\mu(abcd)}{\mu(bc)^2}(\mu(bc) - \mu(bcd))(\mu(bc) - \mu(abc)) \\
&= \sigma^2_{abcd} \,;
\end{aligned}$$

la troisième égalité découle de $\mu(abcd) = \dfrac{\mu(abc)\mu(bcd)}{\mu(bc)}$.

• Cas d'un mot du type $abbc$: en utilisant le Lemme 3, on a

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(N_{abbc} \mid \mathcal{S}_2) = \mu(abbc) - \frac{\mu(abbc)^2}{\mu(abc)}(2d_{abc}+1) + \mu(abbc)^2\left(\sum_{x,y}\frac{d^2_{xy+}}{\mu(xy)} - \sum_{x,y,z}\frac{d^2_{xyz}}{\mu(xyz)}\right).$$

Or, seuls d_{abb}, d_{bbc} et d_{abc} sont non nuls ($d_{abb} = d_{bbc} = 1$ et $d_{abc} = -1$) ; par conséquent, seul d_{bb+} est non nul (égal à 1) car $d_{ab+} = d_{abb} + d_{abc} = 0$. Ainsi,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(N_{abbc} \mid \mathcal{S}_2) &= \mu(abbc) + \frac{\mu(abbc)^2}{\mu(abc)} + \mu(abbc)^2\left(\frac{1}{\mu(bb)} - \frac{1}{\mu(abb)} - \frac{1}{\mu(bbc)} - \frac{1}{\mu(abc)}\right) \\
&= \frac{\mu(abbc)}{\mu(bb)}\left(\mu(bb) + \mu(abbc) - \frac{\mu(abbc)\mu(bb)}{\mu(abb)} - \frac{\mu(abbc)\mu(bb)}{\mu(bbc)}\right) \\
&= \frac{\mu(abbc)}{\mu(bb)}\left(\mu(bb) + \frac{\mu(abb)\mu(bbc)}{\mu(bb)} - \mu(bbc) - \mu(abb)\right) \\
&= \frac{\mu(abbc)}{\mu(bb)^2}(\mu(bb) - \mu(bbc))(\mu(bb) - \mu(abb)) \\
&= \sigma^2_{abbc}\ ;
\end{aligned}$$

la troisième égalité découle de $\mu(abbc) = \dfrac{\mu(abb)\mu(bbc)}{\mu(bb)}$.

- Cas d'un mot du type $aaaa$: d'après le Lemme 3

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(N_{aaaa} \mid \mathcal{S}_2) &= \mu(aaaa) + 2\mu(aaaaa) - \frac{\mu(aaaa)^2}{\mu(aaa)}(2d_{aaa} + 1) \\
&\quad + \mu(aaaa)^2\left(\sum_{x,y} \frac{d^2_{xy+}}{\mu(xy)} - \sum_{x,y,z} \frac{d^2_{xyz}}{\mu(xyz)}\right).
\end{aligned}$$

Or, seul d_{aaa} est non nul ($d_{aaa} = n_{aaa} - 1 = 1$) ; d'où, seul d_{aa+} est non nul de valeur 1. Ainsi,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(N_{aaaa} \mid \mathcal{S}_2) &= \mu(aaaa) + 2\mu(aaaaa) + \mu(aaaa)^2\left(\frac{1}{\mu(aa)} - \frac{1}{\mu(aaa)}\right) - 3\,\frac{\mu(aaaa)^2}{\mu(aaa)} \\
&= \frac{\mu(aaaa)}{\mu(aa)}\left(\mu(aa) + 2\,\frac{\mu(aa)\mu(aaaaa)}{\mu(aaaa)} + \mu(aaaa) - 4\,\frac{\mu(aaaa)\mu(aa)}{\mu(aaa)}\right) \\
&= \frac{\mu(aaaa)}{\mu(aa)}\left(\mu(aa) - 2\mu(aaa) + \frac{\mu(aaa)^2}{\mu(aa)}\right) \\
&= \frac{\mu(aaaa)}{\mu(aa)^2}(\mu(aa) - \mu(aaa))^2 \\
&= \sigma^2_{aaaa}\ ;
\end{aligned}$$

la troisième égalité découle de $\mu(aaaaa) = \dfrac{\mu(aaaa)\mu(aaa)}{\mu(aa)}$ et $\mu(aaaa) = \dfrac{\mu(aaa)^2}{\mu(aa)}$. □

A.3 Moment conditionnel d'ordre 4 dans M2

Lemme 5 : Moment conditionnel d'ordre 4

$$\mathbb{E}\left((N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))^4 \mid \mathcal{S}_2\right) = O(n^2)\ p.s$$

Preuve : On se placera ici dans le cas où UV n'est pas un mot compatible (U et V étant respectivement les première et dernière lettres de $\mathbb{W}$) ; dans le cas compatible, la démonstration est identique à celle de Prum *et al.* (1995).

Par la formule du binôme, on a

$$\begin{aligned}
\mathbb{E}\left((N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))^4 \mid \mathcal{S}_2\right) &= \mathbb{E}\left((N_{\mathbb{W}} - \mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}))^4 \mid \mathbb{S}\right) \\
&= \mathbb{E}(N_{\mathbb{W}}^4 \mid \mathbb{S}) - 4\,\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S})\mathbb{E}(N_{\mathbb{W}}^3 \mid \mathbb{S}) + 6\,\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S})^2\mathbb{E}(N_{\mathbb{W}}^2 \mid \mathbb{S}) \\
&\quad -3\,\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S})^4,
\end{aligned}$$

puis on exprime les moments conditionnels de $N_{\mathbb{W}}$ en fonction de R_k (Prum *et al.*, 1995) :

$$\begin{aligned}
\mathbb{E}(N_{\mathbb{W}} \mid \mathbb{S}) &= n_1 R_1 \\
\mathbb{E}(N_{\mathbb{W}}^2 \mid \mathbb{S}) &= n_1 R_1 + n_2(n_2-1)R_2 + T_2 \\
\mathbb{E}(N_{\mathbb{W}}^3 \mid \mathbb{S}) &= n_1 R_1 + 3n_2(n_2-1)R_2 + n_3(n_3-1)(n_3-2)R_3 + T_3 \\
\mathbb{E}(N_{\mathbb{W}}^4 \mid \mathbb{S}) &= n_1 R_1 + 7n_2(n_2-1)R_2 + 6n_3(n_3-1)(n_3-2)R_3 \\
&\quad + n_4(n_4-1)(n_4-2)(n_4-3)R_4 + T_4
\end{aligned}$$

où $n_1 = N_{UV}(\mathbb{S}\backslash\mathbb{W}) = 1$, $n_2 = N_{UV}(\mathbb{S}\backslash\mathbb{W}\backslash\mathbb{W}) = 2\ \ldots$, et les termes T_2, T_3 et T_4 sont ceux qui tiennent compte des chevauchements éventuels ($T_i = \mathrm{O}(n^{i-1})$).

On fait un développement asymptotique de $\mathbb{E}\left((N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))^4 \mid \mathcal{S}_2\right)$ en s'intéressant plus particulièrement aux termes d'ordre 3 et 4. On traitera séparément les termes faisant apparaître les chevauchements, que l'on désignera par T, c'est à dire $T = T_4 - 4R_1T_3 + 6R_1^2T_2$.

D'après le Lemme 1, presque sûrement

$$R_k = \frac{(N_{U+})^k}{k!}C^k\left(1 + \frac{\Delta_k}{n} + \mathrm{o}(\frac{1}{n})\right),$$

donc

$$\begin{aligned}
\mathbb{E}\left((N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))^4 \mid \mathcal{S}_2\right) &= (N_{U+})^4C^4 + 6(N_{U+})^3C^3 + C^4\Delta_4\frac{(N_{U+})^4}{n} - 4(N_{U+})^4C^4 \\
&\quad -4\,\frac{(N_{U+})^4}{n}C^4(\Delta_1+\Delta_3) - 12(N_{U+})^3C^3 + 6(N_{U+})^4C^4 \\
&\quad + 6(N_{U+})^3C^3 + 6\frac{(N_{U+})^4}{n}C^4(\Delta_2 + 2\Delta_1) - 3(N_{U+})^4C^4 \\
&\quad -12C^4\Delta_1\frac{(N_{U+})^4}{n} + T + \mathrm{O}(n^2) \\
&= \frac{(N_{U+})^4}{n}C^4(\Delta_4 - 4\Delta_1 - 4\Delta_3 + 6\Delta_2) + T + \mathrm{O}(n^2) \\
&= \frac{(N_{U+})^4}{n}C^4(\Delta_4^* - 4\Delta_1^* - 4\Delta_3^* + 6\Delta_2^*) + T + \mathrm{O}(n^2)
\end{aligned}$$

car $(\Delta_4^{**} - 4\Delta_1^{**} - 4\Delta_3^{**} + 6\Delta_2^{**}) = \Delta_1^{**}(4 - 4 - 12 + 12) = 0$.
Or $\Delta_4^* - 4\Delta_1^* - 4\Delta_3^* + 6\Delta_2^*$ est une somme de termes du type

$$4d_{AB}(4d_{AB} - 1) - 4d_{AB}(d_{AB} - 1) - 12d_{AB}(3d_{AB} - 1) + 12d_{AB}(2d_{AB} - 1) = 0 \, .$$

Par conséquent

$$\mathbb{E}\left((N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))^4 \mid \mathcal{S}_2\right) = T + \mathrm{O}(n^2) \quad p.s.$$

Le terme T n'a pas de termes d'ordre 4 ; vérifions donc que les termes d'ordre 3 disparaissent. On peut écrire, presque sûrement

$$\begin{aligned}
T_2 &= \underbrace{2\, N_{U+} \sum_{d=1}^{h-3} \delta(\mathbb{W}; d) C_d}_{a_1} + \mathrm{O}(1) \\
T_3 &= \underbrace{6\, N_{U+}^2\, C \sum_{d=1}^{h-3} \delta(\mathbb{W}; d) C_d}_{a_2} + \mathrm{O}(n) \\
T_4 &= \underbrace{12\, N_{U+}^3\, C^2 \sum_{d=1}^{h-3} \delta(\mathbb{W}; d) C_d}_{a_3} + \mathrm{O}(n^2)
\end{aligned}$$

où C est défini par le Lemme 1, et C_d est l'analogue de C pour le mot chevauché $\mathbb{W}^d\mathbb{W}$.
Par conséquent,

$$T = a_3 - 4\, a_2 C N_{U+} + 6\, a_1 C^2 N_{U+}^2 + \mathrm{O}(n^2)\, ,$$

et plus précisément

$$T = (12 - 4 * 6 + 6 * 2)\, N_{U+}^3 C^2 \sum_{d=1}^{h-3} \delta(\mathbb{W}; d) C_d \; + \; \mathrm{O}(n^2) = \mathrm{O}(n^2)\, ,$$

ce qui démontre le lemme. □

A.4 Moment d'ordre 4 dans M2

Pour démontrer le Théorème 1, en section 2.3, nous avons tout d'abord prouvé la normalité asymptotique suivante :

$$T_W = \frac{1}{\sqrt{n}} \left(N_W - \mathbb{E}(N_W \mid \mathcal{S}_2)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)\, ,$$

puis nous avons montré la convergence presque sûre de $\mathrm{Var}(T_W \mid \mathcal{S}_2)$ vers une variable v^2. Enfin, pour montrer que $\sigma^2 = v^2$, nous avons vérifié que le moment d'ordre 4 de T_W était borné, en calculant le moment conditionnel d'ordre 4 de T_W (Lemme 5).

Suite à une remarque faite par Jean Bretagnolle, rapporteur de cette thèse, on peut tout à fait, sur ce dernier point, faire l'économie du calcul du moment conditionnel d'ordre 4. En effet, il suffit de montrer que $\mathbb{E}(N_W - \mathbb{E}N_W)^4 = \mathrm{O}(n^2)$ (Lemme 6), et il en découle alors que $\mathbb{E}(T_W^4) = \mathrm{O}(1)$ (Corollaire 3).

Lemme 6 *Il existe une constante C ne dépendant que de la chaîne telle que*

$$\mathbb{E}(N_W - \mathbb{E}N_W)^4 \leq C\, n^2.$$

Preuve: On a une chaîne de Markov $(X_i)_{i=-\infty}^{i=+\infty}$ sur un espace d'états fini, apériodique et irréductible. Quitte à changer d'alphabet, on peut se placer dans le cas d'une chaîne d'ordre 1 ; soit Π sa matrice de transition. On munit la chaîne de sa probabilité invariante μ. Un théorème ergodique donne alors l'existence de deux constantes positives c et ρ, avec $\rho < 1$ telles que

$$d_{\mathrm{VT}}(\Pi^n(x,\cdot) - \mu(\cdot)) \leq c\,\rho^n$$

uniformément en x ; d_{VT} désigne la distance en variation totale.
Autrement dit, si $\mathcal{F}_n$ est la σ-algèbre du passé $\sigma(X_i \,|\, i \leq n)$ et $\mathcal{F}^n$ celle de l'avenir $\sigma(X_i \,|\, i \geq n)$, on a la condition de φ-mélange suivante,

$$\text{si } B \in \mathcal{F}^{k+n},\ |\mathrm{P}\{B \mid \mathcal{F}^k\} - \mathrm{P}\{B\}| \leq c\,\rho^n = \varphi'(n).$$

Si l'on considère maintenant le processus stationnaire $Y_i = \mathbb{1}\{X_i \cdots X_{i+h-1} = W\}$, alors la condition de φ-mélange ci-dessus est encore vérifiée avec toutefois $\varphi(n) = c\,\rho^{n-h}$, car Y_i fait intervenir les variables $X_i, \ldots, X_{i+h-1}$. On peut alors utiliser le résultat suivant (Billingsley, 1968, p195) :

Lemme 7 *Si ξ_i est un processus stationnaire, centré, borné en valeur absolue par 1 avec probabilité 1, et φ-mélangeant, si $D = \sum (k+1)^2 \sqrt{\varphi(k)} < \infty$, alors*

$$\mathbb{E}\left((\sum_{i=1}^{n} \xi_i)^4 \right) \leq 288\left(n^2 \mathbb{E}^2(\xi_1^2) + nD^2\mathbb{E}(\xi_1^2)\right).$$

Par conséquent, le Lemme 6 est démontré en posant $\xi_i = Y_i - \mathbb{E}Y_i$, où $\mathbb{E}Y_i = \mu(W)$. □

Corollaire 3 : Moment d'ordre 4

$$\mathbb{E}(T_W^4) = O(1).$$

Preuve: On écrit $N_W - \mathbb{E}(N_W \mid \mathcal{S}_2) = (N_W - \mathbb{E}N_W) + (\mathbb{E}N_W - \mathbb{E}(N_W \mid \mathcal{S}_2))$ et on utilise l'inégalité $(a+b)^4 \leq 8(a^4+b^4)$, a, $b > 0$. Ainsi,

$$\mathbb{E}(T_W^4) \leq 8\left(\mathbb{E}\left(\frac{N_W - \mathbb{E}N_W}{\sqrt{n}}\right)^4 + \mathbb{E}\left(\frac{\mathbb{E}N_W - \mathbb{E}(N_W \mid \mathcal{S}_2)}{\sqrt{n}}\right)^4 \right).$$

De plus, $\mathbb{E}N_W - \mathbb{E}(N_W \mid \mathcal{S}_2) = \mathbb{E}(N_W - \mathbb{E}N_W \mid \mathcal{S}_2)$ et par l'inégalité de Jensen, on a

$$\mathbb{E}\left(\mathbb{E}(N_W - \mathbb{E}N_W \mid \mathcal{S}_2)^4\right) \leq \mathbb{E}(N_W - \mathbb{E}N_W)^4.$$

Par conséquent,

$$\mathbb{E}(T_W^4) \leq 16\, \mathbb{E}\left(\frac{N_W - \mathbb{E}N_W}{\sqrt{n}}\right)^4$$

et le membre de droite est borné d'après le Lemme 6. □

A.5 Espérance et variance conditionnelles dans M0

Nous donnons ici les expressions de l'espérance et la variance conditionnelles, ainsi que leurs formules asymptotiques, du nombre N_W d'occurrences d'un mot W de longueur h, dans le modèle *M0* où les bases $(X_i)_{i=1\cdots n}$ de la séquence sont supposées indépendantes.
Dans ce modèle, la statistique exhaustive par laquelle nous conditionnons est l'ensemble des nombres d'occurrences de A, C, G, et T dans la séquence, et nous la notons $\mathcal{N}$:

$$\mathcal{N} = \{N_{\mathtt{A}},\, N_{\mathtt{C}},\, N_{\mathtt{G}},\, N_{\mathtt{T}}\}\,.$$

Calcul de l'espérance conditionnelle

En utilisant la définition du nombre d'occurrences N_W, d'une part, et l'indépendance des variables X_i, d'autre part, on écrit

$$\begin{aligned}
\mathbb{E}(N_W \mid \mathcal{N}) &= \sum_{i=1}^{n-h+1} \mathrm{P}\{X_i = w_1, \ldots, X_{i+h-1} = w_h \mid \mathcal{N}\} \\
&= (n-h+1)\, \mathrm{P}\{X_1 = w_1, \ldots, X_h = w_h \mid \mathcal{N}\} \\
&= (n-h+1)\, \frac{\mathrm{P}\{X_1 = w_1, \ldots, X_h = w_h,\, N_{\mathtt{A}},\, N_{\mathtt{C}},\, N_{\mathtt{G}},\, N_{\mathtt{T}}\}}{\mathrm{P}\{N_{\mathtt{A}},\, N_{\mathtt{C}},\, N_{\mathtt{G}},\, N_{\mathtt{T}}\}}\,.
\end{aligned}$$

En notant n_a le nombre d'occurrences de la lettre a dans le mot W, et μ_a la probabilité d'apparition de la lettre a dans la séquence, le numérateur vaut

$$\mathrm{P}\{X_1 = w_1, \ldots, X_h = w_h,\, N_{\mathtt{A}},\, N_{\mathtt{C}},\, N_{\mathtt{G}},\, N_{\mathtt{T}}\} = \mu(W) \frac{(n-h)!}{\prod_a (N_a - n_a)!} \prod_a \mu_a^{N_a - n_a}$$

et le dénominateur

$$\mathrm{P}\{N_{\mathtt{A}},\, N_{\mathtt{C}},\, N_{\mathtt{G}},\, N_{\mathtt{T}}\} = \frac{n!}{\prod_a N_a!} \prod_a \mu_a^{N_a}\,.$$

On obtient ainsi, l'expression de l'espérance conditionnelle suivante :

$$\mathbb{E}(N_W \mid \mathcal{N}) = \frac{(n-h+1)!}{n!} \prod_a \frac{N_a!}{(N_a - n_a)!}\,.$$

Pour en faire un développement asymptotique, on utilise le développement limité suivant : $\dfrac{N!}{(N-n)!} = N^n \left(1 - \dfrac{n(n-1)}{2N} + \mathrm{o}\left(\dfrac{1}{N}\right)\right)$. Ainsi, presque sûrement

$$\mathbb{E}(N_W \mid \mathcal{N}) = \frac{\prod_a N_a^{n_a}}{n^{h-1}} \left(1 + \frac{(h-1)(h-2)}{2n} - \sum_a \frac{n_a(n_a-1)}{2N_a} + \mathrm{o}\left(\frac{1}{n}\right)\right),$$

ou encore

$$\mathbb{E}(N_W \mid \mathcal{N}) = n\widehat{\mu}_0(W) + \mathrm{O}(1), \tag{A.6}$$

où $\widehat{\mu}_0(W)$ est l'estimateur naturel de $\mu(W)$ suivant :

$$\widehat{\mu}_0(W) = \prod_{j=1}^{h} \frac{N(w_j)}{n}.$$

Calcul de la variance conditionnelle

En reprenant le développement standard de la variance conditionnelle, en particulier celui de l'espérance conditionnelle du carré de N_W, on obtient

$$\begin{aligned} \mathrm{Var}(N_W \mid \mathcal{N}) &= \mathbb{E}(N_W \mid \mathcal{N}) - (\mathbb{E}(N_W \mid \mathcal{N}))^2 + 2 \sum_{d=1}^{h-1} \mathbb{E}(N_{W^{(d)}W} \mid \mathcal{N}) \qquad \text{(A.7)} \\ &+ 2 \sum_{i=1}^{n-h+1} \sum_{j=i+h}^{n-h+1} \mathrm{P}\{X_i = w_1, \ldots, X_{i+h-1} = w_h, X_j = w_1, \ldots, X_{j+h-1} = w_h \mid \mathcal{N}\}. \end{aligned}$$

Puisque les variables X_i sont indépendantes, la probabilité conditionnelle qui apparaît dans le dernier terme s'écrit aussi

$$\begin{aligned} \mathrm{P}\{X_i = w_1, \ldots, X_{i+h-1} = w_h, X_j = w_1, \ldots, X_{j+h-1} = w_h \mid \mathcal{N}\} &= \mathrm{P}\{X_1 X_2 \cdots X_{2h} = WW \mid \mathcal{N}\} \\ &= \frac{(n-2h)!}{n!} \prod_a \frac{N_a!}{(N_a - 2n_a)!}. \end{aligned}$$

En reportant cette expression dans (A.7), on a

$$\begin{aligned} \mathrm{Var}(N_W \mid \mathcal{N}) &= \mathbb{E}(N_W \mid \mathcal{N}) - (\mathbb{E}(N_W \mid \mathcal{N}))^2 + 2 \sum_{d=1}^{h-1} \mathbb{E}(N_{W^{(d)}W} \mid \mathcal{N}) \\ &+ (n-h+1)(n-3h+2) \frac{(n-2h)!}{n!} \prod_a \frac{N_a!}{(N_a - 2n_a)!}. \end{aligned}$$

En faisant un développement asymptotique presque sûr de chacun des termes, on obtient

$$\begin{aligned} \mathrm{Var}(N_W \mid \mathcal{N}) &= n\mu(W) - n^2\mu^2(W) \left(1 + \frac{2(h-1)(h-2)}{2n} - 2\sum_a \frac{n_a(n_a-1)}{2N_a}\right) \\ &+ 2 \sum_{d=1}^{h-1} n\mu(W^{(d)}W) \\ &+ n^2\mu^2(W) \left(1 + \frac{4h^2 - 10h + 6}{2n} - \sum_a \frac{2n_a(2n_a-1)}{2N_a}\right) + \mathrm{o}(n) ; \end{aligned}$$

on s'aperçoit que les termes d'ordre 2 s'éliminent et il reste

$$\frac{1}{n}\mathrm{Var}(N_W \mid \mathcal{N}) = \mu(W) + 2\sum_{d=1}^{h-1}\mu(W^{(d)}W) + \mu^2(W)\left((h-1)^2 - \sum_a \frac{n_a^2}{\mu(a)}\right) + \mathrm{o}(1). \qquad (A.8)$$

Annexe B

Compléments au Chapitre 3 : le modèle M1_3

B.1 Variance conditionnelle des mots de longueur $h \neq 3p+2$

Le calcul de $\mathrm{Var}(N_{W,k} \mid \mathbb{S})$, pour les mots de longueur $3p+1$ ou $3p$ (Section 3.3.2), fait intervenir des termes de la forme

$$\mathbb{E}(N_{\overline{W'}}\mathbb{I}\{\overline{S} \text{ commence par } \overline{W''}\} \mid \mathbb{S})$$

où W' et W'' sont deux mots de longueur respectives $h' = 3p+2$ et $h'' = 3p+1$. C'est pourquoi, nous calculons ici cette quantité, en remplaçant toutefois $\overline{W''}$ par $\overline{W}$ pour alléger les notations ; de plus on pose $u = w_1$ et $v = w_h$, et on a $h' = h+1$, $\overline{W} = (W, k)$ et $\overline{W'} = (W', k')$. On a

$$\mathbb{E}(N_{\overline{W'}}\mathbb{I}\{\overline{S} \text{ commence par } \overline{W}\} \mid \mathbb{S}) = \sum_{i=h+1}^{n} \mathrm{P}\{\overline{W'} \text{ termine en } i \text{ et } \overline{W} \text{ termine en } h \mid \mathbb{S}\}.$$

On développe cette somme en deux termes :

- un terme $\widetilde{T}_1$ correspondant à $i \geq 2h$. Il correspond au cas où $\overline{W}$ apparaît au début de la séquence et $\overline{W'}$ ne le chevauche pas, ou éventuellement le chevauche sur une lettre.
- Un terme $\widetilde{T}_2$ correspondant à $h+1 \leq i < 2h$. Celui-ci correspond au cas où le mot $\overline{W'}$ chevauche, sur au moins deux lettres, le mot $\overline{W}$ qui commence la séquence.

– Le premier terme est donc

$$\widetilde{T}_1 = \frac{1}{\#\mathbb{S}} \sum_{i=2h}^{n} \beta_{i,h}(\overline{W'}, \overline{W}, \mathbb{S})$$

où $\beta_{i,h}(\overline{W'}, \overline{W}, \mathbb{S})$ est le nombre de séquences dans $\mathbb{S}$ qui commencent par $\overline{W}$ et dans lesquelles le mot $\overline{W'}$ apparaît en position i. Si $\widetilde{\mathbb{S}}^{\overline{W}}$ est l'ensemble des séquences de $\mathbb{S}$ qui commencent par

$\overline{W}$, et $\beta_{i-h+1}(\overline{W'}, \widetilde{\mathbb{S}}^{\overline{W}})$ est le nombre de séquence dans $\widetilde{\mathbb{S}}^{\overline{W}}$ où le mot $\overline{W'}$ apparaît en position $i-h+1$, alors on a

$$\beta_{i,h}(\overline{W'}, \overline{W}, \mathbb{S}) = \beta_{i-h+1}(\overline{W'}, \widetilde{\mathbb{S}}^{\overline{W}}) .$$

Par conséquent,

$$\begin{aligned} \widetilde{T}_1 &= \frac{1}{\#\mathbb{S}} \sum_{i=h+1}^{n-h+1} \beta_i(\overline{W'}, \widetilde{\mathbb{S}}^{\overline{W}}) \\ &= \frac{\#\widetilde{\mathbb{S}}^{\overline{W}}}{\#\mathbb{S}} \mathbb{E}(N_{\overline{W'}} |\, \widetilde{\mathbb{S}}^{\overline{W}}) , \end{aligned}$$

et l'espérance se calcule avec la formule habituelle ; par convention, cette espérance est nulle si $\widetilde{\mathbb{S}}^{\overline{W}} = \emptyset$, c'est à dire si $u \neq \overline{X}_1$ ou si $h\%3 \neq k$.

– Le second terme est

$$\begin{aligned} \widetilde{T}_2 &= \frac{1}{\#\mathbb{S}} \sum_{i=h+1}^{2h-1} \beta_{i,h}(\overline{W'}, \overline{W}, \mathbb{S}) \\ &= \frac{1}{\#\mathbb{S}} \sum_{d=0}^{h-2} \delta(\overline{W}; \overline{W'}; d)\ \#\widetilde{\mathbb{S}}^{\overline{W}^{(d)}\overline{W'}} \end{aligned}$$

où $\widetilde{\mathbb{S}}^{\overline{W}^{(d)}\overline{W'}}$ est l'ensemble des séquences de $\mathbb{S}$ qui commencent par $\overline{W}^{(d)}\overline{W'}$.

On obtient ainsi l'expression suivante

$$\mathbb{E}(N_{\overline{W'}} \mathbb{1}\{\overline{S} \text{ commence par } \overline{W}\} |\, \mathbb{S}) = \frac{\#\widetilde{\mathbb{S}}^{\overline{W}}}{\#\mathbb{S}} \mathbb{E}(N_{\overline{W'}} |\, \widetilde{\mathbb{S}}^{\overline{W}}) + \sum_{d=0}^{h-2} \delta(\overline{W}; \overline{W'}; d)\ \frac{\#\widetilde{\mathbb{S}}^{\overline{W}^{(d)}\overline{W'}}}{\#\mathbb{S}} . \quad \text{(B.1)}$$

B.2 Calcul de Var($N_W \mid$ S) pour $h = 3p+1$ et $h = 3p$

• **W est un mot de longueur** $h = 3p+1$

En utilisant la décomposition

$$N_W = \sum_a N_{aW} + \mathbb{1}\{\overline{S} \text{ commence par } (W,1)\}$$

on a

$$\begin{aligned} \mathrm{Var}(N_W \mid \mathbb{S}) &= \sum_{a,b} \mathrm{Cov}(N_{aW}, N_{bW} \mid \mathbb{S}) \\ &+ \mathrm{P}\{\overline{S} \text{ commence par } (W,1) \mid \mathbb{S}\} \\ &+ 2 \sum_a \mathrm{Cov}(N_{aW}, \mathbb{1}\{\overline{S} \text{ commence par } (W,1)\} \mid \mathbb{S}) . \end{aligned}$$

– Le premier terme s'obtient en écrivant

$$\mathrm{Cov}(N_{aW}, N_{bW} \mid \mathbb{S}) = \mathbb{E}(N_{aW} N_{bW} \mid \mathbb{S}) - \mathbb{E}(N_{aW} \mid \mathbb{S}) \mathbb{E}(N_{bW} \mid \mathbb{S}) ,$$

et

$$\mathbb{E}(N_{aW}N_{bW} \mid \mathbb{S}) = \sum_{k,k'} \mathbb{E}(N_{aW,k}N_{bW,k'} \mid \mathbb{S}),$$

puis en utilisant les formules (3.19) et (3.17).

– Le second, $\mathrm{P}\{\overline{S}$ commence par $(W,1) \mid \mathbb{S}\}$, est donné par (3.12).

– Enfin, le calcul du dernier terme découle de celui de $\mathbb{E}(N_{aW}\,\mathbb{1}\{\overline{S}$ commence par $(W,1)\} \mid \mathbb{S})$:
$\mathbb{E}(N_{aW}\,\mathbb{1}\{\overline{S}$ commence par $(W,1)\} \mid \mathbb{S}) =$

$$\begin{aligned} &= \sum_{i=h+1}^{n} \mathrm{P}\{aW \text{ termine en } i \text{ et } W \text{ termine en } h \mid \mathbb{S}\} \\ &= \frac{\#\widetilde{\mathbb{S}}^{(W)}}{\#\mathbb{S}} \mathbb{E}(N_{aW} \mid \widetilde{\mathbb{S}}^{(W)}) + \sum_{d=0}^{h-2} \delta(W;aW;d)\ \frac{\#\widetilde{\mathbb{S}}^{(W^{(d)}aW)}}{\#\mathbb{S}} \end{aligned} \tag{B.2}$$

où $\widetilde{\mathbb{S}}^{(W)}$ (resp. $\widetilde{\mathbb{S}}^{(W^{(d)}aW)}$) est l'ensemble des séquences de $\mathbb{S}$ qui commencent par W (resp. $W^{(d)}aW$) ; (calcul analogue à celui détaillé en **B.1**).

• **W est un mot de longueur $h = 3p$**

En utilisant la décomposition

$$N_W = \sum_{a,b} N_{abW} + \mathbb{1}\{\overline{S} \text{ commence par } (W,3)\} + \sum_{c} \mathbb{1}\{\overline{S} \text{ commence par } (cW,1)\},$$

on a

$$\begin{aligned} \mathrm{Var}(N_W \mid \mathbb{S}) \;=\; & \sum_{a,b,a',b'} \mathrm{Cov}(N_{abW}, N_{a'b'W} \mid \mathbb{S}) \\ &+ \mathrm{P}\{\overline{S} \text{ commence par } (W,3) \mid \mathbb{S}\} \\ &+ \sum_{c} \mathrm{P}\{\overline{S} \text{ commence par } (cW,1) \mid \mathbb{S}\} \\ &+ 2\sum_{c} \mathrm{P}\{\overline{S} \text{ commence par } (W,3) \text{ et } (cW,1) \mid \mathbb{S}\} \\ &+ 2\sum_{a,b} \mathrm{Cov}(N_{abW}, \mathbb{1}\{\overline{S} \text{ commence par } (W,3)\} \mid \mathbb{S}) \\ &+ 2\sum_{a,b,c} \mathrm{Cov}(N_{abW}, \mathbb{1}\{\overline{S} \text{ commence par } (cW,1)\} \mid \mathbb{S}). \end{aligned}$$

– Le premier terme, ainsi que le dernier, se calculent comme pour le cas précédent ($h = 3p+1$).

– Les termes $\mathrm{P}\{\overline{S}$ commence par $(W,3) \mid \mathbb{S}\}$ et $\mathrm{P}\{\overline{S}$ commence par $(cW,1) \mid \mathbb{S}\}$ sont donnés par (3.14) et (3.12).

– $\mathrm{P}\{\overline{S}$ commence par $(W,3)$ et $(cW,1) \mid \mathbb{S}\} = \delta((W,3);(cW,1);0)$

$$\times\ \mathrm{P}\{\overline{S} \text{ commence par } \underbrace{cc\cdots c}_{h+1} \mid \mathbb{S}\}$$

et il suffit d'appliquer la formule (3.12) avec le mot $(cc\cdots c, 1)$ ($h+1$ lettres).

– Le calcul de $\mathrm{Cov}(N_{abW}, \mathbb{1}\{\overline{S}$ commence par $(W,3)\} \mid \mathbb{S})$ découle de celui de $\mathbb{E}(N_{abW}\, \mathbb{1}\{\overline{S}$ commence par $(W,3)\} \mid \mathbb{S})$; on écrit

$$\mathbb{E}(N_{abW}\, \mathbb{1}\{\overline{S} \text{ commence par } (W,3)\} \mid \mathbb{S}) = \sum_c \mathbb{E}(N_{abW}\, \mathbb{1}\{\overline{S} \text{ commence par } (Wc,1)\} \mid \mathbb{S})$$

et on utilise la formule (B.2) pour calculer le membre de droite.

B.3 Covariance conditionnelle asymptotique

On se propose ici de calculer la limite de $\frac{1}{n}\mathrm{Cov}(N_{W,k}, N_{W',k'} \mid \mathbb{S})$ lorsque W et W' sont deux mots de longueur $h = 3p+2$. W et W' sont de la forme $W = u \dots v$ et $W' = u' \dots v'$, et on notera $\overline{W} = (W,k)$, $\overline{W'} = (W',k')$, $\overline{uv} = (uv,k)$ et $\overline{u'v'} = (u'v',k')$. Il est important de noter que tous les $N_{ab,k}$ tendent vers l'infini quand n tend vers l'infini, car on a supposé que $\pi_k(a,b) > 0$.

Nous introduisons les notations suivantes :

$$\begin{array}{ll} n_0 = N_{\overline{uv}}(\mathbb{S}) & n'_0 = N_{\overline{u'v'}}(\mathbb{S}) \\ d_{\overline{ab}} = N_{\overline{ab}}(\overline{W}) - \mathbb{1}\{\overline{ab} = \overline{uv}\} & d'_{\overline{ab}} = N_{\overline{ab}}(\overline{W'}) - \mathbb{1}\{\overline{ab} = \overline{u'v'}\} \\ n_2 = N_{\overline{uv}}(\mathbb{S} \setminus \overline{W} \setminus \overline{W'}) & n'_2 = N_{\overline{u'v'}}(\mathbb{S} \setminus \overline{W} \setminus \overline{W'}) \ ; \end{array}$$

et, si $\overline{a} = (a, \ell-1)$, $\left\{ \begin{array}{l} d_{\overline{a}+} = \sum_{\overline{b}} d_{\overline{ab}} = \sum_{b,\ell} d_{ab,\ell} \\ d'_{\overline{a}+} = \sum_{\overline{b}} d'_{\overline{ab}} = \sum_{b,\ell} d'_{ab,\ell} \end{array} \right.$.

On a ainsi $n_2 = n_0 - d_{\overline{uv}} - d'_{\overline{uv}}$ et $n'_2 = n'_0 - d_{\overline{u'v'}} - d'_{\overline{u'v'}}$ et on notera que $\frac{n_0}{n} \overset{ps}{\to} \mu(\overline{uv})$ et $\frac{n'_0}{n} \overset{ps}{\to} \mu(\overline{u'v'})$ quand n tend vers l'infini.

Dans un premier temps, à l'aide de la formule de Whittle (1955), nous étudions le comportement asymptotique des rapports $(\#\mathbb{S} \setminus \overline{W} \setminus \overline{W'})/(\#\mathbb{S})$, $(\#\mathbb{S} \setminus \overline{W})/(\#\mathbb{S})$ et $(\#\mathbb{S} \setminus \overline{W'})/(\#\mathbb{S})$ intervenant dans l'expression de la covariance conditionnelle (formules (3.10) et (3.17)). Les deux derniers rapports font en fait l'objet du Lemme 1 de Prum *et al.* (1995) ; le lemme ci-dessous traite du premier.
Nous calculons ensuite la limite de la covariance conditionnelle (Lemme 9).

Lemme 8 : Comportement asymptotique de $(\#\mathbb{S} \setminus \overline{W} \setminus \overline{W'})/(\#\mathbb{S})$

Lorsque n tend vers l'infini,

$$\frac{\#\mathbb{S} \setminus \overline{W} \setminus \overline{W'}}{\#\mathbb{S}} = CC' \left(1 + \frac{\Delta}{n} + o(\frac{1}{n})\right) \quad p.s.$$

avec

$$C = \prod_{\substack{a,b,\ell \\ (ab,\ell) = \overline{ab}}} \left(\frac{N_{\overline{ab}}}{N_{a+,\ell}}\right)^{d_{\overline{ab}}} \quad \overset{ps}{\longrightarrow} \quad \prod_{\substack{a,b,\ell \\ (ab,\ell) = \overline{ab}}} \pi_\ell(a,b)^{d_{\overline{ab}}} = \frac{\mu(\overline{W})}{\mu(\overline{uv})}$$

$$C' = \prod_{\substack{a,b,\ell \\ (ab,\ell) = \overline{ab}}} \left(\frac{N_{\overline{ab}}}{N_{a+,\ell}}\right)^{d'_{\overline{ab}}} \quad \overset{ps}{\longrightarrow} \quad \prod_{\substack{a,b,\ell \\ (ab,\ell) = \overline{ab}}} \pi_\ell(a,b)^{d'_{\overline{ab}}} = \frac{\mu(\overline{W'})}{\mu(\overline{u'v'})}$$

et

$$\Delta = \Delta^* + \Delta^{**} ,$$

où Δ^* *est obtenu en développant* $K(\mathbb{S} \setminus \overline{W} \setminus \overline{W'})/K(\mathbb{S})$,

$$\Delta^* = \sum_{\overline{a}} \frac{(d_{\overline{a}+} + d'_{\overline{a}+})(d_{\overline{a}+} + d'_{\overline{a}+} - 1)}{2} \frac{n}{N_{\overline{a}+}} - \sum_{\overline{ab}} \frac{(d_{\overline{ab}} + d'_{\overline{ab}})(d_{\overline{ab}} + d'_{\overline{ab}} - 1)}{2} \frac{n}{N_{\overline{ab}}} ,$$

Δ^{**} *est obtenu en développant* $H(\mathbb{S} \setminus \overline{W} \setminus \overline{W'})/H(\mathbb{S})$ *et satisfait*

$$\frac{n_0 n'_0}{n^2}\Delta^{**} = \frac{n'_0}{n}\Delta^{**}_W + \frac{n_0}{n}\Delta^{**}_{W'} ,$$

Δ^{**}_W *(resp.* $\Delta^{**}_{W'}$*) étant obtenu en développant* $H(\mathbb{S} \setminus \overline{W})/H(\mathbb{S})$ *(resp.* $H(\mathbb{S} \setminus \overline{W'})/H(\mathbb{S})$*) (voir Prum et al., 1995).*

Preuve: En reprenant la définition de K et en utilisant le fait que $N_{\overline{ab}}(\mathbb{S} \setminus \overline{W} \setminus \overline{W'}) = N_{\overline{ab}} - d_{\overline{ab}} - d'_{\overline{ab}}$, on a

$$\frac{K(\mathbb{S} \setminus \overline{W} \setminus \overline{W'})}{K(\mathbb{S})} = \prod_{\overline{a}} \frac{(N_{\overline{a}+} - d_{\overline{a}+} - d'_{\overline{a}+})!}{N_{\overline{a}+}!} \times \prod_{\overline{ab}} \frac{N_{\overline{ab}}!}{(N_{\overline{ab}} - d_{\overline{ab}} - d'_{\overline{ab}})!} .$$

Or,

$$\frac{N!}{(N-d)!} = N^d \left(1 - \frac{d(d-1)}{2N} + \mathrm{o}(\frac{1}{N})\right) \text{ si } d \geq 0 \text{ et } N \text{ grand} ;$$

d'où, si $d_{\overline{ab}} + d'_{\overline{ab}} \geq 0 \quad \forall\, \overline{ab}$, on a presque sûrement

$$\begin{aligned}
\frac{K(\mathbb{S} \setminus \overline{W} \setminus \overline{W'})}{K(\mathbb{S})} &= \prod_{\overline{a}} \left[N_{\overline{a}+}^{d_{\overline{a}+}+d'_{\overline{a}+}} \left(1 - \frac{(d_{\overline{a}+} + d'_{\overline{a}+})(d_{\overline{a}+} + d'_{\overline{a}+} - 1)}{2N_{\overline{a}+}} + \mathrm{o}(\frac{1}{N_{\overline{a}+}})\right)\right]^{-1} \\
&\quad \times \prod_{\overline{ab}} N_{\overline{ab}}^{d_{\overline{ab}}+d'_{\overline{ab}}} \left(1 - \frac{(d_{\overline{ab}} + d'_{\overline{ab}})(d_{\overline{ab}} + d'_{\overline{ab}} - 1)}{2N_{\overline{ab}}} + \mathrm{o}(\frac{1}{N_{\overline{ab}}})\right) \\
&= \prod_{\overline{a}=(a,\ell)} \frac{\prod_{\{b | \overline{ab}=(ab,\ell+1)\}} N_{\overline{ab}}^{d_{\overline{ab}}+d'_{\overline{ab}}} \left(1 - \frac{(d_{\overline{ab}}+d'_{\overline{ab}})(d_{\overline{ab}}+d'_{\overline{ab}}-1)}{2N_{\overline{ab}}} + \mathrm{o}(\frac{1}{N_{\overline{ab}}})\right)}{N_{\overline{a}+}^{d_{\overline{a}+}+d'_{\overline{a}+}} \left(1 - \frac{(d_{\overline{a}+}+d'_{\overline{a}+})(d_{\overline{a}+}+d'_{\overline{a}+}-1)}{2N_{\overline{a}+}} + \mathrm{o}(\frac{1}{N_{\overline{a}+}})\right)} \\
&= \prod_{\overline{a}=(a,\ell)} \frac{\prod_{\{b | \overline{ab}=(ab,\ell+1)\}} N_{\overline{ab}}^{d_{\overline{ab}}+d'_{\overline{ab}}}}{N_{\overline{a}+}^{d_{\overline{a}+}+d'_{\overline{a}+}}} \left(1 + \frac{\Delta^*}{n} + \mathrm{o}(\frac{1}{n})\right) \\
&= \prod_{\substack{a,b,\ell \\ (ab,\ell) = \overline{ab}}} \left(\frac{N_{\overline{ab}}}{N_{a+,\ell}}\right)^{d_{\overline{ab}}+d'_{\overline{ab}}} \left(1 + \frac{\Delta^*}{n} + \mathrm{o}(\frac{1}{n})\right) .
\end{aligned}$$

Si $\exists\ \overline{ab}$ tel que $d_{\overline{ab}} + d'_{\overline{ab}} < 0$ (c'est à dire égal à -1 ou -2), on fait le même raisonnement et on arrive à la même formule asymptotique.

Quant à $H(\mathbb{S}\setminus\overline{W}\setminus\overline{W'})/H(\mathbb{S})$, on remarque que, d'une part $H(\mathbb{S})$ est une forme polynômiale en les paramètres $\widehat{\pi}_k(a,b)$ et que, d'autre part

$$(\widehat{\pi}_k(a,b)\,|\,\mathbb{S}\setminus\overline{W}\setminus\overline{W'}) = \widehat{\pi}_k(a,b)\left(1+\frac{D_{\overline{ab}}}{n}+\mathrm{o}(\frac{1}{n})\right)\,,$$

avec

$$D_{\overline{ab}} = \frac{n}{N_{\overline{a}+}}(d_{\overline{a}+}+d'_{\overline{a}+}) - \frac{n}{N_{\overline{ab}}}(d_{\overline{ab}}+d'_{\overline{ab}})\,.$$

Par conséquent, on peut écrire

$$\frac{H(\mathbb{S}\setminus\overline{W}\setminus\overline{W'})}{H(\mathbb{S})} = 1+\frac{\Delta^{**}}{n}+\mathrm{o}(\frac{1}{n}) \quad p.s.\,;$$

en remarquant que

$$D_{\overline{ab}} = \frac{n}{n_0}\left(\frac{n_0}{N_{\overline{a}+}}d_{\overline{a}+}-\frac{n_0}{N_{\overline{ab}}}d_{\overline{ab}}\right)+\frac{n}{n'_0}\left(\frac{n'_0}{N_{\overline{a}+}}d'_{\overline{a}+}-\frac{n'_0}{N_{\overline{ab}}}d'_{\overline{ab}}\right)$$

il est clair que $\frac{n_0 n'_0}{n^2}\Delta^{**} = \frac{n'_0}{n}\Delta^{**}_W + \frac{n_0}{n}\Delta^{**}_{W'}$, où Δ^{**}_W et $\Delta^{**}_{W'}$ sont définis dans Prum *et al.* (1995). Ce qui démontre le lemme. $\square$

En utilisant le lemme précédent et le Lemme 1 de Prum *et al.* (1995), on obtient le résultat suivant.

Lemme 9 : Covariance conditionnelle asymptotique dans *M1_3*

$$\begin{aligned}\lim_{n\to+\infty}\frac{1}{n}Cov(N_{\overline{W}},N_{\overline{W'}}\,|\,\mathbb{S}) &= \sum_{d=2-h}^{h-2}\delta(\overline{W};\overline{W'};d)\,\mu(\overline{W}^{(d)}\overline{W'})\\ &- \frac{\mu(\overline{W})\,\mu(\overline{W'})}{\mu(\overline{uv})\,\mu(\overline{u'v'})}\left(\mu(\overline{uv})\,d_{\overline{u'v'}}+\mu(\overline{u'v'})\,d'_{\overline{uv}}+\mu(\overline{uv})\ \mathbb{1}\{\overline{uv}=\overline{u'v'}\}\right)\\ &+ \mu(\overline{W})\,\mu(\overline{W'})\left(\sum_{\overline{a}}\frac{d_{\overline{a}+}\,d'_{\overline{a}+}}{\mu(\overline{a})}-\sum_{\overline{ab}}\frac{d_{\overline{ab}}\,d'_{\overline{ab}}}{\mu(\overline{ab})}\right)\quad p.s.\end{aligned}$$

Preuve : D'après la formule (3.10) et le Lemme 1 de Prum *et al.* (1995), on a

$$\begin{aligned}\mathbb{E}(N_{\overline{W}}\,|\,\mathbb{S}) &= N_{\overline{uv}}(\mathbb{S}\setminus\overline{W})\,\frac{\#\mathbb{S}\setminus\overline{W}}{\#\mathbb{S}}\\ &= (n_0-d_{\overline{uv}})\,C\left(1+\frac{\Delta_W}{n_0}+\mathrm{o}(\frac{1}{n_0})\right)\,,\end{aligned}$$

avec

$$\Delta_W = \Delta^*_W + \Delta^{**}_W$$

et

$$\Delta^*_W = \sum_{\overline{a}} \frac{d_{\overline{a}+}(d_{\overline{a}+}-1)}{2} \, \frac{n_0}{N_{\overline{a}+}} - \sum_{\overline{ab}} \frac{d_{\overline{ab}}(d_{\overline{ab}}-1)}{2} \, \frac{n_0}{N_{\overline{ab}}} \,.$$

Par le Lemme 8, la formule (3.17) devient

$$\begin{aligned}
\mathbb{E}(N_{\overline{W}} N_{\overline{W'}} \,|\, \mathbb{S}) &= \sum_{d=2-h}^{h-2} \delta(\overline{W};\overline{W'};d)\mathbb{E}(N_{\overline{W}^{(d)}\overline{W'}} \,|\, \mathbb{S}) \\
&+ (n_0 - d_{\overline{uv}} - d'_{\overline{uv}})\,(n'_0 - d_{\overline{u'v'}} - d'_{\overline{u'v'}} - \mathbb{1}\{\overline{uv} = \overline{u'v'}\})CC'\left(1 + \frac{\Delta}{n} + \mathrm{o}(\frac{1}{n})\right)
\end{aligned}$$

et par conséquent, le développement asymptotique de la covariance devient

$$\begin{aligned}
\frac{1}{n}\mathrm{Cov}(N_{\overline{W}}, N_{\overline{W'}} \,|\, \mathbb{S}) &= \sum_{d=2-h}^{h-2} \delta(\overline{W};\overline{W'};d)\,\frac{1}{n}\mathbb{E}(N_{\overline{W}^{(d)}\overline{W'}} \,|\, \mathbb{S}) \\
&+ CC'\left(\left(\frac{n_0 n'_0}{n^2}\Delta - \frac{n_0}{n}\Delta_{W'} - \frac{n'_0}{n}\Delta_W\right) - \frac{n_0}{n}(d_{\overline{u'v'}} + \mathbb{1}\{\overline{uv} = \overline{u'v'}\}) - \frac{n'_0}{n}d'_{\overline{uv}}\right) \\
&+ \mathrm{O}(\frac{1}{n})
\end{aligned}$$

Or, d'après la relation qui lie Δ^{**}, Δ^{**}_W et $\Delta^{**}_{W'}$, on a

$$\frac{n_0 n'_0}{n^2}\Delta - \frac{n_0}{n}\Delta_{W'} - \frac{n'_0}{n}\Delta_W = \frac{n_0 n'_0}{n^2}\Delta^* - \frac{n_0}{n}\Delta^*_{W'} - \frac{n'_0}{n}\Delta^*_W \,.$$

En remplaçant Δ^*, Δ^*_W et $\Delta^*_{W'}$ par leur expression, on obtient

$$\frac{n_0 n'_0}{n^2}\Delta - \frac{n_0}{n}\Delta_{W'} - \frac{n'_0}{n}\Delta_W = \sum_{\overline{a}} d_{\overline{a}+}\, d'_{\overline{a}+}\, \frac{n_0 n'_0}{n N_{\overline{a}+}} - \sum_{\overline{ab}} d_{\overline{ab}}\, d'_{\overline{ab}}\, \frac{n_0 n'_0}{n N_{\overline{ab}}} \,.$$

Enfin, on a les convergences presque sûres suivante quand n tend vers l'infini

$$\frac{n_0}{n} \xrightarrow{ps} \mu(\overline{uv}) \,, \quad \frac{n'_0}{n} \xrightarrow{ps} \mu(\overline{u'v'}) \,, \quad \frac{N_{\overline{a}+}}{n} \xrightarrow{ps} \mu(\overline{a}) \,, \quad \frac{N_{\overline{ab}}}{n} \xrightarrow{ps} \mu(\overline{ab}) \,,$$

$$\text{et} \quad \frac{1}{n}\mathbb{E}(N_{\overline{W}^{(d)}\overline{W'}} \,|\, \mathbb{S}) \xrightarrow{ps} \mu(\overline{W}^{(d)}\overline{W'}) \,,$$

ce qui démontre le lemme. □

Annexe C

Distance entre deux lois de Poisson ou deux lois de Poisson composées

Le Théorème 13 donne une majoration de la distance en variation totale entre la loi du nombre $\widetilde{N}^*(W)$ de trains d'un mot W et la loi de Poisson d'espérance $n\widetilde{\mu}(W)$. L'objet de cette annexe est de vérifier que l'on peut approcher la loi de $\widetilde{N}^*(W)$ par la loi de Poisson d'espérance $n\widehat{\widetilde{\mu}}(W)$, où $\widehat{\widetilde{\mu}}(W)$ est un estimateur consistant de $\widetilde{\mu}(W)$ dans le modèle *M1* . Pour cela, en utilisant l'inégalité triangulaire, on calcule la distance en variation totale entre les deux lois de Poisson d'espérances $n\widetilde{\mu}(W)$ et $n\widehat{\widetilde{\mu}}(W)$, et on s'assure qu'elle tend vers zéro lorsque n augmente. De la même façon, on cherchera à majorer la distance en variation totale entre deux lois de Poisson composées pour voir ce que devient l'inégalité donnée par le Théorème 14, lorsque l'on remplace les paramètres de la loi de Poisson composée $\mathrm{CP}(\Lambda)$ par des estimateurs consistants.

Le théorème ci-dessous est un résultat classique qui majore la distance en variation totale entre deux lois de Poisson par l'écart absolu entre leurs espérances respectives. Il est suivi d'une démonstration basée sur l'approximation d'une somme de variables aléatoires de Bernoulli indépendantes et identiquement distribuées, par une loi de Poisson.

Théorème 16 *Soient U et V deux variables aléatoires de Poisson d'espérances respectives μ et ν. La distance en variation totale d_{VT} entre les lois $\mathcal{L}(U)$ de U et $\mathcal{L}(V)$ de V satisfait*

$$d_{VT}(\mathcal{L}(U), \mathcal{L}(V)) \leq |\mu - \nu|.$$

Preuve: On peut approcher U par $U_r = U_{r1} + U_{r2} + \cdots + U_{rr}$ où $U_{r1}, \ldots, U_{rr}$ sont des variables aléatoires de Bernoulli indépendantes d'espérances $\frac{\mu}{r}$. De même, on peut approcher V par $V_r = V_{r1} + V_{r2} + \cdots + V_{rr}$ où $V_{r1}, \ldots, V_{rr}$ sont des variables aléatoires de Bernoulli indépendantes d'espérances $\frac{\nu}{r}$, et indépendantes des variables aléatoires U_{ri} pour $i = 1 \cdots r$.

Par l'inégalité triangulaire, on a :

$$d_{\mathrm{VT}}(\mathcal{L}(U), \mathcal{L}(V)) \leq d_{\mathrm{VT}}(\mathcal{L}(U), \mathcal{L}(U_r)) + d_{\mathrm{VT}}(\mathcal{L}(U_r), \mathcal{L}(V_r)) + d_{\mathrm{VT}}(\mathcal{L}(V_r), \mathcal{L}(V)).$$

On utilise alors les inégalités

$$\begin{aligned} d_{\mathrm{VT}}(\mathcal{L}(U), \mathcal{L}(U_r)) &\leq \frac{\mu}{r}(1 - \mathrm{e}^{-\mu}) \leq \frac{\mu}{r}, \\ d_{\mathrm{VT}}(\mathcal{L}(V), \mathcal{L}(V_r)) &\leq \frac{\nu}{r}(1 - \mathrm{e}^{-\nu}) \leq \frac{\nu}{r}, \end{aligned}$$

dues à Barbour et Hall (1984). De plus

$$d_{\mathrm{VT}}(\mathcal{L}(U_r), \mathcal{L}(V_r)) \leq r\, d_{\mathrm{VT}}(\mathcal{L}(U_{r1}), \mathcal{L}(V_{r1}))$$

car les v.a. U_{ri} sont i.i.d. ainsi que les v.a. V_{ri}. Or,

$$\begin{aligned} d_{\mathrm{VT}}(\mathcal{L}(U_{r1}), \mathcal{L}(V_{r1})) &= \max(\mathrm{P}\{U_{r1} = 0\} - \mathrm{P}\{V_{r1} = 0\},\ \mathrm{P}\{U_{r1} = 1\} - \mathrm{P}\{V_{r1} = 1\}) \\ &= \max\left(\frac{\nu - \mu}{r}, \frac{\mu - \nu}{r}\right) \\ &= \frac{|\mu - \nu|}{r}. \end{aligned}$$

Par conséquent,

$$d_{\mathrm{VT}}(\mathcal{L}(U), \mathcal{L}(V)) \leq \frac{\mu}{r} + |\mu - \nu| + \frac{\nu}{r}.$$

En faisant tendre r vers l'infini, on obtient le résultat souhaité. □

Le Théorème 16 permet de quantifier l'erreur introduite par l'estimation du paramètre lorsque l'on approche la loi du nombre $\widetilde{N}^*(W)$ de trains de W par la loi de Poisson d'espérance $n\widehat{\widetilde{\mu}}(W)$; cette erreur est donc de l'ordre de $\mathrm{O}(n|\widehat{\widetilde{\mu}}(W) - \widetilde{\mu}(W)|)$. Afin de vérifier que cette erreur tend vers zéro lorsque n tend vers l'infini, on étudie dans un premier temps le cas d'un mot W qui ne se chevauche pas, ce qui se traduit par $\widetilde{N}^*(W) = N(W)$ et $\widetilde{\mu}(W) = \mu(W)$. Par l'inégalité du triangle, on a

$$d_{\mathrm{VT}}(\mathcal{L}(N(W)), \mathcal{P}\mathrm{o}(n\widehat{\mu}(W))) \leq d_{\mathrm{VT}}(\mathcal{L}(N(W)), \mathcal{P}\mathrm{o}(n\mu(W))) + n|\widehat{\mu}(W) - \mu(W)|$$

où $\widehat{\mu}(W)$ est un estimateur consistant de $\mu(W)$. Le Théorème 13 donne une majoration du premier terme, tandis que les deux propositions suivantes donnent l'ordre de grandeur de l'écart $n|\widehat{\mu}(W) - \mu(W)|$ pour l'estimateur

$$\widehat{\mu}(W) = \frac{N(w_1)}{n} \prod_{a,b \in \mathcal{A}} \left(\frac{N(ab)}{N(a+)}\right)^{n_{ab}} \tag{C.1}$$

où w_1 est la première lettre de W et n_{ab} est le nombre de 2-mots ab contenus dans W.

Proposition 6 *Soit W un mot de longueur h. Alors, on a*

$$\widehat{\mu}(W) = \mu(W)\left(1 + O\left(\frac{h\sqrt{\log\log n}}{\sqrt{n}}\right)\right) \quad p.s.$$

Preuve : En utilisant la loi du log itéré pour les chaînes de Markov à espace d'états fini, on a

$$\frac{N(w_1)}{n} - \mu(w_1) = \mathrm{O}\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right) \quad p.s.$$

où μ est la mesure invariante de la chaîne (Meyn et Tweedie, 1993, Théorème 17.0.1), et

$$\frac{N(ab)}{N(a+)} - \pi(a,b) = \mathrm{O}\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right) \quad p.s.$$

où $\pi(a,b)$ est la probabilité de transition de a vers b de la chaîne de Markov (Senoussi (1990)). Par conséquent,

$$\widehat{\mu}(W) = \left(\mu(w_1) + \mathrm{O}\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right)\right) \prod_{a,b} \left(\pi(a,b) + \mathrm{O}\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right)\right)^{n_{ab}}.$$

Puisque les transitions $\pi(a,b)$ sont strictement positives (hypothèse du modèle) et en nombre fini, on a presque sûrement

$$\begin{aligned}
\widehat{\mu}(W) &= \mu(w_1) \prod_{a,b} (\pi(a,b))^{n_{ab}} \left(1 + \mathrm{O}\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right)\right) \left(1 + \sum_{a,b} n_{ab} \mathrm{O}\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right)\right) \\
&= \mu(W) \left(1 + \mathrm{O}\left(\frac{h\sqrt{\log\log n}}{\sqrt{n}}\right)\right) \ ;
\end{aligned}$$

la dernière égalité utilise la relation $h = \sum_{a,b} n_{ab}$. Ceci démontre ainsi la proposition. □

La Proposition 7 découle immédiatement de la Proposition 6.

Proposition 7 *Soit W un mot de longueur h ;*

$$\begin{aligned}
&(i) \quad si \quad n\mu(W) = O(1), \ alors\ n\widehat{\mu}(W) - n\mu(W) = O\left(\frac{h\sqrt{\log\log n}}{\sqrt{n}}\right) \quad p.s., \\
&(ii) \quad si \quad n\mu(W) = O(1) \ et\ h = o(\sqrt{n}), \ alors\ n\widehat{\mu}(W) - n\mu(W) = o(1) \quad p.s.
\end{aligned}$$

Dans le cas général d'un mot éventuellement chevauchant, on reprend les notations du Chapitre 5 et on choisit comme estimateur de $\widetilde{\mu}(W)$

$$\widehat{\widetilde{\mu}}(W) = \widehat{\mu}(W) - \sum_{p \in \mathcal{P}'(W)} \widehat{\mu}(W^{(p)}W),$$

où les estimateurs $\widehat{\mu}(\cdot)$ ont été définis plus haut par (C.1). On a alors le résultat suivant :

Proposition 8 *Soit W un mot de longueur h ;*

$$\begin{aligned}
&(i) \quad si \quad n\mu(W) = O(1), \ alors\ n\widehat{\widetilde{\mu}}(W) - n\widetilde{\mu}(W) = O\left(\frac{h^2\sqrt{\log\log n}}{\sqrt{n}}\right) \quad p.s., \\
&(ii) \quad si \quad n\mu(W) = O(1) \ et\ h = o(n^{1/4}), \ alors\ n\widehat{\widetilde{\mu}}(W) - n\widetilde{\mu}(W) = o(1) \quad p.s.
\end{aligned}$$

Preuve : Puisque $\widetilde{\mu}(W) = \mu(W) - \sum_{p\in\mathcal{P}'(W)} \mu(W^{(p)}W)$, la Proposition 7 donne

$$n\widehat{\widetilde{\mu}}(W) = n\widetilde{\mu}(W) + \mathrm{O}\left(\frac{h\sqrt{\log\log n}}{\sqrt{n}}\right) - \sum_{p\in\mathcal{P}'(W)} \mathrm{O}\left(\frac{(h+p)\sqrt{\log\log n}}{\sqrt{n}}\right) \quad p.s.$$

Or, $1 \le p \le h$ et la somme porte sur au plus h termes. Ainsi,

$$n\widehat{\widetilde{\mu}}(W) = n\widetilde{\mu}(W) + \mathrm{O}\left(\frac{h^2\sqrt{\log\log n}}{\sqrt{n}}\right) \quad p.s.,$$

ce qui démontre (i) ; (ii) en découle immédiatement. □

À partir des résultats pour les lois de Poisson, on peut majorer la distance en variation totale entre deux lois de Poisson composée.

Soient $\mathbb{X} = (X_k)_{k\geq 1}$ une suite de v.a. indépendantes et distribuées suivant les lois de Poisson d'espérances μ_k, et $\mathbb{Y} = (Y_k)_{k\geq 1}$ une suite de v.a. indépendantes et distribuées suivant les lois de Poisson d'espérances ν_k. Les lois de Poisson composées de paramètres respectifs $(\mu_k)_{k\geq 1}$ et $(\nu_k)_{k\geq 1}$ sont les lois des variables $X = \sum_{k\geq 1} k\,X_k$ et $Y = \sum_{k\geq 1} k\,Y_k$. On a ainsi

$$\begin{aligned} d_{\mathrm{VT}}(\mathcal{L}(X), \mathcal{L}(Y)) &\leq d_{\mathrm{VT}}(\mathcal{L}(\mathbb{X}), \mathcal{L}(\mathbb{Y})) \\ &\leq \sum_{k\geq 1} d_{\mathrm{VT}}(\mathcal{L}(X_k), \mathcal{L}(Y_k)). \end{aligned}$$

Il en résulte, par le Théorème 16, que

$$d_{\mathrm{VT}}(\mathcal{L}(X), \mathcal{L}(Y)) \leq \sum_{k\geq 1} |\mu_k - \nu_k|. \tag{C.2}$$

Par conséquent, si $\mathrm{CP}(\Lambda)$ est la loi de Poisson composée définie au Théorème 14 avec $\Lambda = \sum_{k\geq 1} n\widetilde{\mu}_k\,\delta_k$, et $\mathrm{CP}(\widehat{\Lambda})$ est la loi de Poisson composée définie par $\widehat{\Lambda} = \sum_{k\geq 1} n\widehat{\widetilde{\mu}_k}\,\delta_k$, où $\widehat{\widetilde{\mu}_k}$ est un estimateur de $\widetilde{\mu}_k$, on a

$$d_{\mathrm{VT}}(\mathrm{CP}(\widehat{\Lambda}), \mathrm{CP}(\Lambda)) \leq \sum_{k\geq 1} n|\widehat{\widetilde{\mu}_k} - \widetilde{\mu}_k|.$$

Ainsi, si l'on veut approcher la loi de $N(W)$ par la loi de Poisson composée $\mathrm{CP}(\widehat{\Lambda})$, il faut s'assurer que, sous les hypothèses du Corollaire 2, $\sum_{k\geq 1} n|\widehat{\widetilde{\mu}_k} - \widetilde{\mu}_k|$ tend vers zéro lorsque l'on fait tendre n vers l'infini.

Dans un premier temps, nous simplifions l'expression de $\widetilde{\mu}_k$ donnée par la formule (5.8) :

$$\widetilde{\mu}_k = \sum_{C\in\mathcal{C}_k} \mu(C) \;-2 \sum_{C'\in\mathcal{C}_{k+1}} \mu(C') \;+ \sum_{C''\in\mathcal{C}_{k+2}} \mu(C'').$$

En effet, si $C' \in \mathcal{C}_{k+1}$ alors il existe $p \in \mathcal{P}'(W)$ et $C \in \mathcal{C}_k$ tel que $C' = W^{(p)}C$. Plus précisément, $\mathcal{C}_{k+1} = \{W^{(p)}C,\, p \in \mathcal{P}'(W),\, C \in \mathcal{C}_k\}$. Or, si $(w_i)_{i=1\cdots h}$ sont les h lettres du mot W, on a

$$\begin{aligned} \mu(W^{(p)}C) &= \mu(W^{(p)})\pi(w_p, w_1)\frac{\mu(C)}{\mu(w_1)} \\ &= \pi(w_1, w_2)\cdots\pi(w_p, w_{p+1})\mu(C) \end{aligned}$$

car $w_{p+1} = w_1$. Par conséquent, en notant

$$A_n = \sum_{p \in \mathcal{P}'(W)} \prod_{j=1}^{p} \pi(w_j, w_{j+1}),$$

on a

$$\sum_{C' \in \mathcal{C}_{k+1}} \mu(C') = \sum_{C \in \mathcal{C}_k} \sum_{p \in \mathcal{P}'(W)} \mu(W^{(p)}C) = A_n \sum_{C \in \mathcal{C}_k} \mu(C).$$

De la même façon, on obtient

$$\begin{aligned} \sum_{C'' \in \mathcal{C}_{k+2}} \mu(C'') &= \sum_{C' \in \mathcal{C}_{k+1}} \sum_{p \in \mathcal{P}'(W)} \mu(W^{(p)}C') \\ &= A_n^2 \sum_{C \in \mathcal{C}_k} \mu(C). \end{aligned}$$

On obtient ainsi une nouvelle expression de $\widetilde{\mu}_k$:

$$\widetilde{\mu}_k = (1 - A_n)^2 \sum_{C \in \mathcal{C}_k} \mu(C),$$

qui nous permet d'établir la relation de récurrence $\widetilde{\mu}_k = A_n \, \widetilde{\mu}_{k-1}$. Or $\widetilde{\mu}_1 = (1 - A_n)^2 \sum_{C \in \mathcal{C}_1} \mu(C) = (1 - A_n)^2 \mu(W)$, car $\mathcal{C}_1 = \{W\}$.

Par conséquent, on a

$$\widetilde{\mu}_k = (1 - A_n)^2 A_n^{k-1} \mu(W). \tag{C.3}$$

Remarques:

1. A_n ne dépend que du mot W dont la longueur tend vers l'infini avec n.

2. A_n est une quantité positive qui ne s'annule uniquement si le mot W n'est pas chevauchant. Autrement dit, dans le cadre d'une approximation par une loi de Poisson composée, seuls les mots W chevauchants ont un intérêt, et on a donc $A_n > 0$.

3. La relation $\sum_{k \geq 1} k \, \widetilde{\mu}_k = \mu(W)$ démontrée en (5.10), nous garantit de plus que $A_n < 1$.

On choisit maintenant l'estimateur de $\widetilde{\mu}_k$ suivant

$$\widehat{\widetilde{\mu}_k} = (1 - \widehat{A}_n)^2 \widehat{A}_n^{k-1} \widehat{\mu}(W),$$

avec

$$\widehat{A}_n = \sum_{p \in \mathcal{P}'(W)} \prod_{j=1}^{p} \frac{N(w_j w_{j+1})}{N(w_j +)}.$$

En reprenant la démonstration de la formule (C.3), on peut vérifier que cet estimateur $\widehat{\widetilde{\mu}_k}$ est identique à celui qui consiste à remplacer les μ par des $\widehat{\mu}$ dans la formule (5.8) de $\widetilde{\mu}_k$. Par conséquent, $\sum_{k \geq 1} k \, \widehat{\widetilde{\mu}_k} = \widehat{\mu}(W)$ d'où $0 < \widehat{A}_n < 1$.

Pour montrer que

$$\sum_{k\geq 1} |(1-\widehat{A}_n)^2 \widehat{A}_n^{k-1} n\widehat{\mu}(W) - (1-A_n)^2 A_n^{k-1} n\mu(W)| = \mathrm{o}(1) \quad p.s.,$$

on coupe la somme en deux morceaux $\{k \leq K\}$ et $\{k > K\}$ où K est amené à tendre vers l'infini. Ainsi, on majore la somme infinie par une somme de deux termes qui tendent vers 0 de façon géométrique, lorsque K tend vers l'infini. Pour la somme portant sur K termes, il suffit alors de faire un développement limité de $(1-\widehat{A}_n)^2 \widehat{A}_n^{k-1} n\widehat{\mu}(W)$ autour de $(1-A_n)^2 A_n^{k-1} n\mu(W)$, avec

$$\widehat{A}_n = A_n + \mathrm{O}\left(\frac{h^2\sqrt{\log\log n}}{\sqrt{n}}\right).$$

Il faut toutefois considérer trois cas : $A_n = \mathrm{o}\left(\frac{h^2\sqrt{\log\log n}}{\sqrt{n}}\right)$, $1 - A_n = \mathrm{o}\left(\frac{h^2\sqrt{\log\log n}}{\sqrt{n}}\right)$ puis le cas intermédiaire.
En prenant $K = \sqrt{h}$, par exemple, on montre que la somme tend vers 0 pourvu que $h = \mathrm{o}(n^{1/6})$.

Annexe D

Grandes déviations pour le comptage d'un mot dans une chaîne de Markov

L'utilisation des techniques de grandes déviations va nous permettre ici d'obtenir une approximation fine de la probabilité que la fréquence d'un mot dans une chaîne de Markov, sur un alphabet fini, dévie largement de sa fréquence moyenne.

Considérons une chaîne de Markov X_1, X_2, …, X_n stationnaire d'ordre 1 sur un alphabet $\mathcal{A}$ fini, et un mot $W = w_1 w_2 \cdots w_h$ de longueur h sur cet alphabet. En notant N_W le nombre d'occurrences du mot W dans la suite $X_1 X_2 \cdots X_n$, on décrit ci-dessous une méthode d'approximation des probabilités $\mathrm{P}\{\frac{1}{n-h+1} N_W > \mu(W) + B\}$ et $\mathrm{P}\{\frac{1}{n-h+1} N_W < \mu(W) - b\}$ avec $0 < b < 1$, $0 < B < 1$ et μ est la mesure stationnaire de la chaîne de Markov ; $\mu(W)$ représente la probabilité d'observer un mot W dans la chaîne.
Cette méthode est ainsi adaptée pour évaluer la P-valeur $\mathrm{P}\{N_W > N_W^{\mathrm{obs}}\}$ lorsque le mot W est un mot de fréquence exceptionnelle dans une séquence d'ADN.

Soit $\Pi = (\pi(a,b))_{a,b \in \mathcal{A}}$ la matrice de transition de la chaîne de Markov $S = X_1 X_2 \cdots X_n$, avec

$$\pi(a,b) = \mathrm{P}\{X_i = b \mid X_{i-1} = a\} \ ;$$

on suppose que $\pi(a,b) > 0$ pour tout $a, b \in \mathcal{A}$.
On construit la chaîne de Markov $\mathbb{S}$ d'ordre 1 sur l'alphabet $\mathcal{A}^h$, h fixé, de la façon suivante :

$$\mathbb{S} = \mathbb{X}_1 \mathbb{X}_2 \cdots \mathbb{X}_{n-h+1}\,, \text{ avec } \mathbb{X}_i = X_i X_{i+1} \cdots X_{i+h-1} \in \mathcal{A}^h.$$

La matrice de transition $\mathbb{\Pi} = (\Pi(U,V))_{U,V \in \mathcal{A}^h}$ associée est définie par : soient $U = u_1 u_2 \cdots u_h \in \mathcal{A}^h$ et $V = v_1 v_2 \cdots v_h \in \mathcal{A}^h$, alors

$$\begin{aligned}
\Pi(U,V) &= \mathrm{P}\{\mathbb{X}_i = V \mid \mathbb{X}_{i-1} = U\} \\
&= \mathrm{P}\{X_i X_{i+1} \cdots X_{i+h-1} = v_1 v_2 \cdots v_h \mid X_{i-1} X_i \cdots X_{i+h-2} = u_1 u_2 \cdots u_h\} \\
&= \begin{cases} \pi(u_h, v_h) & \text{si } u_{j+1} = v_j,\ j = 1 \cdots h-1, \\ 0 & \text{sinon.} \end{cases}
\end{aligned}$$

Il est clair que beaucoup de coefficients de $\mathbb{\Pi}$ sont nuls mais la matrice $\mathbb{\Pi}^h$ a tous ses coefficients strictement positifs : $\mathbb{S}$ est irréductible sur un un espace d'états fini.

On peut donc appliquer le Théorème de Miller (Bucklew, 1990, p.79) à

$$N_W = \sum_{i=1}^{n-h+1} \mathbb{1}\{\mathbb{X}_i = W\} = \sum_{i=1}^{n-h+1} f(\mathbb{X}_i),$$

ce qui donne l'énoncé suivant.

La loi de $\frac{1}{n-h+1} \sum_{i=1}^{n-h+1} f(\mathbb{X}_i)$ *satisfait un Principe de Grandes Déviations Fort de bonne fonction de taux*

$$I(x) = \sup_{\theta \geq 0} \{\theta x - \log \lambda(\theta)\},$$

où $\lambda(\theta)$ *est la valeur propre de Perron-Frobenius (c'est à dire la plus grande) de la matrice*

$$\mathbb{\Pi}_\theta = \left(e^{\theta f(V)} \Pi(U,V)\right)_{U,V \in \mathcal{A}^h}.$$

Remarque : si $V \neq W$ alors $\mathbb{\Pi}_\theta(U,V) = \mathbb{\Pi}(U,V)$, sinon $\mathbb{\Pi}_\theta(U,V) = \mathrm{e}^\theta \mathbb{\Pi}(U,V)$.

Autrement dit, on a les deux inégalités ci-dessous :

$$\forall F \subset \mathbb{R} \text{ fermé}, \limsup_{n \to +\infty} \frac{1}{n-h+1} \log \mathrm{P}\left\{\frac{1}{n-h+1} N_W \in F\right\} \leq - \inf_{x \in F} I(x),$$

$$\forall O \subset \mathbb{R} \text{ ouvert}, \liminf_{n \to +\infty} \frac{1}{n-h+1} \log \mathrm{P}\left\{\frac{1}{n-h+1} N_W \in O\right\} \geq - \inf_{x \in O} I(x).$$

Or, la fonction de taux I est une fonction positive et convexe, qui s'annule uniquement en la fréquence moyenne $\mu(W)$. Par conséquent, en prenant $F = [\mu(W) + B, 1]$ et $O =]\mu(W) + B, 1[$ avec $B > 0$, on a

$$\mathrm{P}\left\{\frac{1}{n-h+1} N_W \geq \mu(W) + B\right\} \simeq \exp(-(n-h+1) I(\mu(W) + B)) \; ;$$

on a aussi

$$\mathrm{P}\left\{\frac{1}{n-h+1} N_W \leq \mu(W) - B\right\} \simeq \exp(-(n-h+1) I(\mu(W) - B)).$$

Par conséquent,

$$\text{si } N_W^{\text{obs}} > (n-h+1)\mu(W)\,, \qquad \mathrm{P}\{N_W \geq N_W^{\text{obs}}\} \simeq \exp\left(-(n-h+1) I\left(\frac{N_W^{\text{obs}}}{n-h+1}\right)\right)$$

$$\text{si } N_W^{\text{obs}} < (n-h+1)\mu(W)\,, \qquad \mathrm{P}\{N_W \leq N_W^{\text{obs}}\} \simeq \exp\left(-(n-h+1) I\left(\frac{N_W^{\text{obs}}}{n-h+1}\right)\right).$$

Bibliographie

ALTSCHUL, S. F. and ERICKSON, B. W. (1985). Significance of nucleotide sequence alignments : A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.* **2** 526–538.

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1989). Two moments suffice for Poisson approximations : the Chen-Stein method. *Ann. Prob.* **17** 9–25.

ARRATIA, R., GORDON, L. and WATERMAN, M. S. (1990a). The Erdös-Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18** 539–570.

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1990b). Poisson approximation and the Chen-Stein method. *Statistical Science.* **5** 403–434.

AVERY, P. J. (1987). The analysis of intron data and their use in the detection of short signals. *J. Mol. Evol.* **26** 335–340.

BARBOUR, A. D., CHEN, L. H. Y. and LOH, W.-L. (1992a). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20** 1843–1866.

BARBOUR, A. D., HOLST, L. and JANSON, S. (1992b). *Poisson approximation.* Oxford-University Press.

BARBOUR, A. D. and HALL, P. (1984). On the rate of poisson convergence. *Math. Proc. Camb. Phil. Soc.* **95** 473–480.

BILLINGSLEY, P. (1968). *Convergence of probability measures.* Wiley.

BLAISDELL, B. E. (1985). Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucariotic nuclear DNA sequences both protein-coding and noncoding. *J. Mol. Evol.* **21** 278–288.

BRENDEL, V., BECKMANN, J. S. and TRIFONOV, E. N. (1986). Linguistics of nucleotide sequences : Morphology and comparison of vocabularies. *J. Biomol. Struct. Dynamics.* **4** 11–21.

BUCKLEW, J. A. (1990). *Large deviation techniques in decision, simulation, and estimation.* Wiley.

BURGE, C., CAMPBELL, A. and KARLIN, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA.* **89** 1358–1362.

CHEN, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Prob.* **3** 534–545.

CHRYSSAPHINOU, O. and PAPASTAVRIDIS, S. (1988a). A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials. *J. Appl. Prob.* **25** 428–431.

CHRYSSAPHINOU, O. and PAPASTAVRIDIS, S. (1988b). A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials. *Prob. Theory Rel. Fields.* **79** 129–143.

COWAN, R. (1991). Expected frequencies of DNA patterns using Whittle's formula. *J. Appl. Prob.* **28** 886–892.

DACUNHA-CASTELLE, D. and DUFLO, M. (1983). *Probabilités et statistiques 2.Problèmes à temps mobile.* Masson.

DUFLO, M. (1990). *Méthodes récursives aléatoires.* Masson.

FU, J. C. (1993). Poisson convergence in reliability of a large linearly connected system as related to coin tossing. *Statistica Sinica.* **3** 261–275.

GESKE, M. X., GODBOLE, A. P., SCHAFFNER, A. A., SKOLNICK, A. M. and WALLSTROM, G. L. (1995). Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* **32** 877–892.

GÉLIS, F. and SCHBATH, S. (1995). *RMES: Recherche de Mots Exceptionnels dans les Séquences d'ADN – Version 1.* Notice d'utilisation. INRA, Biométrie, F78352 Jouy-en-Josas.

GODBOLE, A. P. (1991). Poisson approximations for runs and patterns of rare events. *Adv. Appl. Prob.* **23** 851–865.

GODBOLE, A. P. and SCHAFFNER, A. A. (1993). Improved poisson approximations for word patterns. *Adv. Appl. Prob.* **25** 334–347.

GUIBAS, L. J. and ODLYZKO, A. M. (1981a). Periods in strings. *J. Combinatorial Theory A.* **30** 19–42.

GUIBAS, L. J. and ODLYZKO, A. M. (1981b). String overlaps, pattern matching, and nontransitive games. *J. Combinatorial Theory A.* **30** 183–208.

Hirano, K. and Aki, S. (1993). On number of occurrences of sucess runs of specified length in a two-state Markov chain. *Statistica Sinica.* **3** 313–320.

Karlin, S., Burge, C. and Campbell, A. M. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl. Acids Res.* **20** 1363–1370.

Karlin, S. and Ost, F. (1987). Counts of long aligned word matches among random letter sequences. *Ann. Prob.* **19** 293–351.

Lothaire, M. (1983). *Combinatorics on words.* Addison-Wesley.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability.* Springer-Verlag.

Nussinov, R. (1981). The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice. *J. Mol. Evol.* **17** 237–244.

Pevzner, P. A., Borodovsky, M. Y. and Mironov, A. A. (1989). Linguistics of nucleotides sequences I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct. Dynamics.* **6** 1013–1026.

Phillips, G. J., Arnold, J. and Ivarie, R. (1987). Mono- through hexanucleotide composition of the *escherichia coli* genome : a Markov chain analysis. *Nucl. Acids Res.* **15** 2611–2626.

Prum, B., Rodolphe, F. and Turckheim, E. d. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

Rajarshi, M. B. (1974). Success runs in a two-state Markov chain. *J. Appl. Prob.* **11** 190–192.

Roos, M. (1994). Stein's method for compound Poisson approximation : the local approach. *Ann. Appl. Prob.* **4** 1177–1187.

Senoussi, R. (1990). Statistique asymptotique presque-sûre de modèles statistiques convexes. *Ann. Inst. Henri Poincaré.* **26** 19–44.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. 6th Berkeley Sympos. Math. Statist. Probab.* **2** 583–602.

Stűckle, E. E., Emmrich, C., Grob, U. and Nielsen, P. J. (1990). Statistical analysis of nucleotide sequences. *Nucl. Acids Res.* **18** 6641–6647.

Whittle, P. (1955). Some distribution and moment formulae for the Markov chain. *J. R. Statist. Soc. B.* **17** 235–242.