6

Séquences exceptionnelles dans l'ADN

Les enzymes de restriction sont des enzymes (endonucléases) capables de couper l'ADN à l'endroit où apparaît une séquence précise, appelée site de restriction. Les enzymes de restriction sont très largement utilisées en biologie moléculaire, par exemple pour le fractionnement de l'ADN, pour la préparation de fragments d'ADN en vue de leur insertion dans l'ADN d'un organisme ou pour la recherche de mutations dans l'ADN. Plusieurs centaines d'enzymes de restriction, avec la séquence du site de restriction associée, sont actuellement répertoriées (voir le site REBASE `http://rebase.neb.com`).

Les enzymes de restriction participent à la défense des bactéries contre les infections virales. En effet, si un virus possède dans son ADN un ou plusieurs sites de restriction, alors, dès qu'il pénètre dans la bactérie, son ADN est découpé par les enzymes de restriction correspondantes de la bactérie. Ensuite, d'autres enzymes interviennent pour dégrader complètement les fragments d'ADN du virus. Il est clair que ce mécanisme de défense peut également se retourner contre la bactérie elle-même. Même si des mécanismes de réparation de l'ADN de la bactérie permettent de compenser les dégradations dues à la présence de ces sites de restriction, on s'attend à ce que les sites de restriction associés aux enzymes de restriction de la bactérie soient des séquences très peu fréquentes dans l'ADN de la bactérie. Voici trois exemples de sites de restriction de Escherichia coli (E. coli) : `GGTCTC`, `CGGCCG`, `CCGCGG` correspondant aux enzymes de restriction Eco 31 I, Eco 52 I et Eco 55 I.

Il existe également des enzymes (exonucléases) qui dégradent l'ADN à partir d'une extrémité du brin d'ADN. Très schématiquement, quand certaines nucléases rencontrent, lors de la dégradation de l'ADN, un motif particulier, appelé motif Chi (acronyme de « cross-over hotspot instigator »), la dégradation s'arrête et un mécanisme de réparation entre alors en jeu (voir [6] pour E. coli). Ce mécanisme permet à la bactérie de se protéger contre ses propres mécanismes de dégradations des ADN étrangers. En particulier, on s'attend à ce que le motif Chi soit une séquence très fréquente dans l'ADN de la bactérie. Pour E. coli, le motif Chi est la séquence `GCTGGTGG`. D'autres

motifs Chi ont également été identifiés dans d'autres micro-organismes (voir par exemple [5]) comme la séquence GCGCGTG pour le Lactococcus lactis.

On comprend à partir de ces deux exemples que certains mots, i.e. certaines courtes séquences de bases, sont pour des raisons biologiques très rares ou très fréquentes. Pour exhiber des mots susceptibles d'avoir une signification biologique, il est intéressant de détecter les mots exceptionnels : des mots très fréquents ou très rares. Après avoir proposé un modèle de chaîne de Markov pour la séquence d'ADN, on peut alors préciser si le nombre d'occurrences d'un mot donné correspond aux prédictions du modèle ou si, au contraire, ce mot est exceptionnel. Des logiciels de recherche de mots exceptionnels ont été développés à partir de tels modèles par le groupe de recherche SSB (Statistiques des Séquences Biologiques) et sont disponibles sur le site de l'INRA (http://www-mig.jouy.inra.fr/ssb/rmes/).

Les paragraphes qui suivent présentent plusieurs approches pour détecter les mots exceptionnels. Ils reposent sur des travaux effectués depuis les années 1990. Nous renvoyons à l'article de Prum, Rodolphe et de Turckheim [8], ainsi qu'aux thèses de Schbath [11] et Nuel [7], pour un exposé rigoureux et plus complet des méthodes utilisées. On pourra également consulter l'ouvrage de Robin, Rodolphe et Schbath, [9], sur le sujet.

Un test d'indépendance pour les paires de bases consécutives de l'ADN met en évidence que l'on ne peut modéliser les séquences de l'ADN comme la réalisation de variables aléatoires à valeurs dans $E = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, indépendantes et de même loi. Dans ce qui suit, nous choisissons donc un modèle plus complexe mais élémentaire de chaîne de Markov : nous supposons que la séquence de l'ADN, de longueur N, $y_1 \dots y_N$, est la réalisation des N premiers termes d'une chaîne de Markov $Y = (Y_n, n \geq 1)$ à valeurs dans E. On note P sa matrice de transition (inconnue). Nous observons, sur une séquence d'ADN assez longue, toutes les successions possibles de paires. Cela implique que pour tous $y, y' \in E$, $P(y, y') > 0$ et donc que $P(y, y') \in]0, 1[$.

Dans le paragraphe 6.1, nous calculons d'abord, grâce au théorème ergodique, le nombre d'occurrences théorique moyen d'un mot (i.e. d'une séquence donnée) pour un modèle général, ainsi que les fluctuations attendues autour de cette moyenne théorique. Puis nous présentons une méthode statistique, autrement dit un test, pour exhiber les mots exceptionnels par rapport aux prédictions du modèle. Il s'agit de mots dont le trop petit ou trop grand nombre d'occurrences n'est pas expliqué par le hasard, tel qu'il est modélisé.

Dans le paragraphe 6.2, nous présentons une variante, qui permet de s'affranchir d'un défaut du test établi dans le paragraphe 6.1 (voir la remarque 6.1.3, point 3).

Les paragraphes 6.1 et 6.2 reposent sur l'analyse des fluctuations associées aux théorèmes ergodiques. Cette analyse est valide quand la séquence d'ADN est très longue devant le mot étudié. Or dès que l'on considère des mots de quelques lettres (par exemple le motif Chi de E. coli), le nombre théorique d'occurrences du mot est faible (quelques unités ou quelques centaines) et l'utilisation du théorème central limite (TCL) n'est plus valide. De fait nous

présentons dans le paragraphe 6.3 des tests reposant sur les « lois des petits nombres » qui sont plus adaptés aux ordres de grandeurs observés en analyse de l'ADN.

Enfin le paragraphe 6.4 présente sous forme de problème une généralisation des résultats du paragraphe 6.1 pour les modèles de chaînes de Markov d'ordre supérieur.

Les théorèmes principaux des paragraphes 6.2 et 6.3 seront admis (voir [8] et [11] pour un exposé complet).

6.1 Fluctuations du nombre d'occurrences d'un mot

Soit une chaîne de Markov $Y = (Y_n, n \geq 1)$ à valeurs dans un espace fini E non réduit à un singleton, de matrice de transition P, telle que pour tous $y, y' \in E$, $P(y, y') \in]0, 1[$. En particulier la chaîne de Markov est irréductible et, d'après la remarque 1.5.7, elle possède une unique probabilité invariante π et $\pi(y) > 0$ pour tout $y \in E$. On utilise la notation y_k^l pour le vecteur $(y_k, \ldots, y_l)$ (avec $k \leq l$).

On appelle mot de longueur $h \geq 1$ une séquence $v = v_1 \cdots v_h$, où $v_i \in E$ pour $1 \leq i \leq h$. Le mot v est également identifié au vecteur $(v_1, \ldots, v_h)$. On note N_v le nombre d'occurrences du mot v dans une séquence, $Y_1, \ldots, Y_N$, de longueur N :

$$N_v = \sum_{k=h}^{N} \mathbf{1}_{\{Y_{k-h+1}^k = v\}}.$$

(Ainsi pour la séquence *abaaa*, on a $N_{ab} = 1$ et $N_{aa} = 2$.) Pour $h \geq 2$, on définit $\pi(v)$ la probabilité en régime stationnaire pour que Y_1^h soit égal au mot v :

$$\pi(v) = \pi(v_1) \prod_{i=1}^{h-1} P(v_i, v_{i+1}). \tag{6.1}$$

Soit $w = w_1 \cdots w_h$ un mot de longueur $h \geq 3$. D'après le théorème ergodique, et plus précisément le corollaire 1.5.11 avec $g(y_1^h) = \mathbf{1}_{\{y_1^h = w\}}$, nous avons l'asymptotique suivante pour le nombre d'occurrences N_w du mot w :

$$\frac{1}{N} N_w \xrightarrow[N \to \infty]{p.s.} \pi(w). \tag{6.2}$$

Nous voulons étudier les fluctuations de N_w par rapport à $N\pi(w)$. Comme $\pi(w)$ n'est pas connu, il faut en donner un estimateur. Remarquons que, si $w- = w_1 \cdots w_{h-1}$ désigne le mot w privé de sa dernière lettre, on a

$$\pi(w) = \pi(w-)P(w_{h-1}, w_h) = \frac{\pi(w-)\pi(w_{h-1}w_h)}{\pi(w_{h-1})}.$$

Le théorème ergodique (corollaire 1.5.11) permet alors de donner un estimateur convergent de $\pi(w)$ sous la forme de $\frac{1}{N}\frac{N_{w-}N_{w_{h-1}w_h}}{N_{w_{h-1}}}$, Il est alors naturel d'étudier la différence entre les deux estimateurs de $\pi(w)$: $\frac{N_w}{N}$ et $\frac{N_{w-}N_{w_{h-1}w_h}}{NN_{w_{h-1}}}$. Dans ce but, on pose

$$\zeta_N = \frac{1}{\sqrt{N}}\left(N_w - \frac{N_{w-}N_{w_{h-1}w_h}}{N_{w_{h-1}}}\right). \tag{6.3}$$

Théorème 6.1.1. *La suite $(\zeta_N, N \geq h)$ converge en loi vers G de loi gaussienne centrée de variance*

$$\sigma^2 = \pi(w)\Big[1 - \frac{\pi(w-)}{\pi(w_{h-1})}\Big][1 - P(w_{h-1}, w_h)].$$

La démonstration complète de ce théorème est reportée à la fin de ce paragraphe.

On définit également

$$\hat{\sigma}_N^2 = \frac{N_w}{N}\left[1 - \frac{N_{w-}}{N_{w_{h-1}}}\right]\left[1 - \frac{N_{w_{h-1}w_h}}{N_{w_{h-1}}}\right] \quad \text{et} \quad \hat{\sigma}_N = \sqrt{\hat{\sigma}_N^2}.$$

Remarquons que le théorème ergodique (corollaire 1.5.11) implique que la suite $(\hat{\sigma}_N^2, N \geq h)$ converge p.s. vers σ^2. Comme $\sigma^2 > 0$, on déduit alors du théorème de Slutsky (théorème A.3.12) le corollaire suivant.

Corollaire 6.1.2. *La suite $Z = (Z_N = \zeta_N/\hat{\sigma}_N, N \geq h)$ converge en loi vers G de loi gaussienne centrée réduite.*

L'exemple 6.2.5 donne une illustration de ce corollaire au travers d'une simulation.

Nous indiquons maintenant comment ce dernier résultat asymptotique permet d'établir une procédure de test pour identifier les mots ou séquences d'ADN exceptionnels. La procédure de test est présentée ici au travers de la notion de p-valeur (voir par exemple [3] pour un traité de statistique).

On note Z_N^{obs} la valeur de Z_N calculée sur la séquence observée $y_1 \cdots y_N$ (par exemple la séquence d'ADN). Plus précisément on remplace les nombres d'occurrences des mots, N_v, par les nombres d'occurrences de ces mêmes mots, N_v^{obs}, observés sur la séquence d'ADN considérée. On s'intéresse ensuite à la p-valeur, p_N^{obs}, associée :

$$p_N^{\text{obs}} = \mathbb{P}(Z_N > Z_N^{\text{obs}}).$$

Remarquons que $p_N^{\text{obs}} = 1 - F_N(Z_N^{\text{obs}})$, où F_N est la fonction de répartition de Z_N. La fonction de répartition F_N n'est pas connue explicitement, on ne peut donc pas calculer la p-valeur. Comme $(Z_N, N \geq h)$ converge en loi vers

G de loi gaussienne centrée réduite de fonction de répartition continue F, on déduit de la proposition C.2, que $(F_N, N \geq h)$ converge simplement vers F. (En fait la convergence est uniforme par le théorème de Dini.) La fonction de répartition, F, de la loi gaussienne, est tabulée (ou programmée), on peut donc calculer numériquement la p-valeur approchée $\tilde{p}_N^{\text{obs}} = 1 - F(Z_N^{\text{obs}})$.

Si le modèle est correct, alors les nombres d'occurrences observés correspondent à des réalisations de variables aléatoires décrites par le modèle. En particulier, la p-valeur approchée du mot w, $\tilde{p}_N^{\text{obs}}$, est une réalisation de la variable aléatoire $\tilde{p}_N = 1 - F(Z_N)$. En particulier, le corollaire 6.1.2 assure que la p-valeur approchée $(\tilde{p}_N, N \geq h)$ converge en loi vers $1 - F(Z)$, où Z est une variable aléatoire gaussienne centrée. Remarquons que F est la fonction de répartition de Z. La proposition C.5 assure que la loi de $F(Z)$, et donc de $1 - F(Z)$, est la loi uniforme sur $[0, 1]$.

Ainsi, la p-valeur approchée $\tilde{p}_N^{\text{obs}}$ est asymptotiquement la réalisation d'une variable aléatoire uniforme. En conclusion, on obtient le test suivant pour détecter si un mot est exceptionnel :

- Si la p-valeur $\tilde{p}_N^{\text{obs}}$ est anormalement faible (proche de 0), cela signifie que la valeur de Z_N^{obs} est anormalement élevée. Cela traduit le fait que N_w^{obs} est anormalement plus grand que $N_{w^-}^{\text{obs}} N_{w_{h-1}w_h}^{\text{obs}} / N_{w_{h-1}}^{\text{obs}}$. On dira alors que le mot w est exceptionnellement fréquent.
- Si la p-valeur $\tilde{p}_N^{\text{obs}}$ est anormalement élevée (proche de 1), cela signifie que N_w^{obs} est anormalement plus petit que $N_{w^-}^{\text{obs}} N_{w_{h-1}w_h}^{\text{obs}} / N_{w_{h-1}}^{\text{obs}}$. On dira alors que le mot w est exceptionnellement rare.

Remarque 6.1.3. Les trois remarques suivantes permettent d'appréhender les limites de cette approche.

- En pratique, on calcule la p-valeur approchée pour tous les mots d'une longueur donnée, et on exhibe ceux dont les p-valeurs sont très faibles ou très élevées. Mais attention, si l'on regarde seulement des mots de longueur $h = 6$ pour un espace d'état à 4 éléments, alors on dispose de $4^6 = 4\,096$ mots et donc de $4\,096$ p-valeurs. Si les nombres d'occurrences des mots de longueur 6 étaient indépendants (ce qui bien sûr n'est pas le cas), alors on observerait $4\,096$ réalisations de variables uniformes indépendantes. Il serait tout à fait naturel d'observer parmi les $4\,096$ p-valeurs des p-valeurs faibles (de l'ordre de $1/4\,096 \simeq 0.0002$) et des p-valeurs élevées (de l'ordre de 0.9998). Signalons que les p-valeurs extrêmes observées pour l'ADN de E. coli, qui correspondent aux résultats numériques du Tableau 6.1, dépassent très largement ces bornes. En revanche pour des simulations, voir l'exemple 6.2.5, les p-valeurs minimales et maximales sont de cet ordre. Nous retiendrons que le seuil de détection des mots exceptionnels dépend du nombre de mots considérés.
- Le TCL pour les chaînes de Markov permet d'obtenir le comportement asymptotique de la p-valeur approchée $\tilde{p}_N$ quand N est grand.

Toutefois, il ne permet pas d'obtenir la précision de cette approximation. Dans le cas de variables aléatoires indépendantes, une indication de cette précision est donnée, par exemple, par le théorème de Berry-Esséen. De manière générale, on observe que l'approximation du TCL est mauvaise pour les grandes valeurs de $|Z_N|$ ou quand on regarde des phénomènes de faible probabilité. On n'obtient pas alors le bon ordre de grandeur de la p-valeur. Dans ces cas, il est souvent préférable d'utiliser d'autres approches asymptotiques. En particulier, si le mot w est long, alors sa probabilité d'apparition est faible, et on s'intéresse alors à des phénomènes rares. Ce dernier aspect peut être abordé par la théorie des grandes déviations (voir [7]) ainsi que par la « loi des petits nombres ». Cette dernière approche est l'objet du paragraphe 6.3.

– En regardant la démonstration du théorème 6.1.1, on peut remarquer que le choix de ζ_N correspond exactement au cadre du TCL pour les chaînes de Markov. Cet opportunisme mathématique ne doit pas masquer la réalité du test construit dans ce paragraphe. En fait, le test construit dans ce paragraphe affirme que le mot w est exceptionnel si l'écart entre N_w et $N_{w-}N_{w_{h-1}w_h}/N_{w_{h-1}}$ est significatif. La quantité $N_{w-}N_{w_{h-1}w_h}/N_{w_{h-1}}$ représente le nombre de mots w escompté connaissant le nombre d'occurrences du mot $w-$, et les nombres d'occurrences de w_{h-1} et $w_{h-1}w_h$. En particulier, si le mot $w-$ est lui-même exceptionnel (rare ou fréquent), il se peut que, conditionnellement au nombre d'occurrences de $w-$, le mot w ne soit pas exceptionnel. Le test construit dans ce paragraphe permet de détecter en fait les mots w qui sont exceptionnels au vu du nombre d'occurrences du mot $w-$. Dans le paragraphe 6.3, on présente un autre test qui permet de s'affranchir de cet artefact.

◊

Démonstration du théorème 6.1.1. On considère la suite de variables aléatoires $X = (X_n, n \geq h-1)$ à valeurs dans E^{h-1} définie par

$$X_n = (Y_{n-h+2}, \ldots, Y_n) = Y_{n-h+2}^n.$$

D'après le lemme 1.5.10, X est une chaîne de Markov irréductible sur E^{h-1} de matrice de transition définie pour $x = x_1^{h-1}, x' = {x'}_1^{h-1} \in E^{h-1}$, par

$$P^X(x, x') = \mathbf{1}_{\{{x'}_1^{h-2} = x_2^{h-1}\}} P(x_{h-1}, x'_{h-1}),$$

et de probabilité invariante $\pi^X(x) = \pi(x_1) \prod_{i=1}^{h-2} P(x_i, x_{i+1})$.

Le nombre d'occurrences du mot w peut se récrire comme

$$N_w = \sum_{k=h}^{N} g_1(X_{k-1}, X_k),$$

où $g_1(x,x') = \mathbf{1}_{\{x_1^{h-1}=w-,x'_{h-1}=w_h\}}$ (rappelons que $w-$ est le mot w tronqué de sa dernière lettre). La proposition 1.6.3 permet alors de préciser la vitesse de convergence de N_w/N vers $\pi(w)$. Toutefois, pour utiliser la proposition 1.6.3 dans cet exemple, il faut évaluer $\sum_{k=h}^N P^X g_1(X_{k-1})$. On calcule pour $x = x_1^{h-1} \in E^{h-1}$

$$\begin{aligned} P^X g_1(x) &= \sum_{x' \in E^{h-1}} P^X(x,x') g_1(x,x') \\ &= \sum_{x'_1,\ldots,x'_{h-1} \in E} \mathbf{1}_{\{x'^{h-2}_1 = x_2^{h-1}\}} P(x_{h-1}, x'_{h-1}) \mathbf{1}_{\{x_1^{h-1}=w-,x'_{h-1}=w_h\}} \\ &= \mathbf{1}_{\{x_1^{h-1}=w-\}} P(w_{h-1}, w_h). \end{aligned}$$

Ainsi, on obtient

$$\begin{aligned} \sum_{k=h}^N P^X g_1(X_{k-1}) &= \sum_{k=h}^N \mathbf{1}_{\{Y_{k-h+1}^{k-1}=w-\}} P(w_{h-1}, w_h) \\ &= N_{w-} P(w_{h-1}, w_h) - \mathbf{1}_{\{Y_{N-h+2}^N = w-\}} P(w_{h-1}, w_h), \end{aligned}$$

où N_{w-} est le nombre d'occurrences du mot $w-$.

Remarquons que l'on ne connaît pas la quantité $P(w_{h-1}, w_h)$. On cherche donc à l'estimer, à l'aide de $X_{h-1}, \ldots, X_N$. Il est naturel d'estimer $P(w_{h-1}, w_h)$ par $N_{w_{h-1}w_h}/N_{w_{h-1}}$. Le nombre d'occurrences du mot $w_{h-1}w_h$, $N_{w_{h-1}w_h}$, peut s'écrire

$$N_{w_{h-1}w_h} = \sum_{k=h}^N g_2(X_{k-1}, X_k) + \sum_{k=2}^{h-1} \mathbf{1}_{\{Y_{k-1}=w_{h-1}, Y_k = w_h\}},$$

avec $g_2(x,x') = \mathbf{1}_{\{x_{h-1}=w_{h-1}, x'_{h-1}=w_h\}}$. Remarquons que l'on a $P^X g_2(x) = \mathbf{1}_{\{x_{h-1}=w_{h-1}\}} P(w_{h-1}, w_h)$ et

$$\begin{aligned} &\sum_{k=h}^N P^X g_2(X_{k-1}) \\ &\quad = N_{w_{h-1}} P(w_{h-1}, w_h) - \left[\sum_{k=1}^{h-2} \mathbf{1}_{\{Y_k = w_{h-1}\}} + \mathbf{1}_{\{Y_N = w_{h-1}\}} \right] P(w_{h-1}, w_h). \end{aligned}$$

Les suites $(\frac{1}{\sqrt{N}} \mathbf{1}_{\{Y_{N-h+2}^N = w-\}}, N \geq h)$, $(\frac{1}{\sqrt{N}} \sum_{k=2}^{h-1} \mathbf{1}_{\{Y_{k-1}=w_{h-1}, Y_k = w_h\}}$, $N \geq h)$ et $(\frac{1}{\sqrt{N}} [\sum_{k=1}^{h-2} \mathbf{1}_{\{Y_k = w_{h-1}\}} + \mathbf{1}_{\{Y_N = w_{h-1}\}}], N \geq h)$ sont positives et majorées par la suite $(h/\sqrt{N}, N \geq h)$. Donc elles convergent p.s. vers 0. On

déduit du théorème de Slutsky et du corollaire 1.6.5, avec la fonction vectorielle $h = (g_1, g_2)$, que la suite $(H_N, N \geq h)$, où

$$H_N = \frac{1}{\sqrt{N}} \begin{pmatrix} N_w - N_{w-} P(w_{h-1}, w_h) \\ N_{w_{h-1}w_h} - N_{w_{h-1}} P(w_{h-1}, w_h) \end{pmatrix},$$

converge en loi vers G un vecteur gaussien centré de matrice de covariance $\Sigma = (\Sigma_{i,j}, 1 \leq i, j \leq 2)$, avec $\Sigma_{i,j} = (\pi^X, P^X(g_i g_j)) - (\pi^X, (P^X g_i)(P^X g_j))$. On explicite ensuite la matrice de covariance. Comme $g_1^2 = g_1$, on obtient

$$\begin{aligned} \Sigma_{11} &= (\pi^X, P^X(g_1)) - (\pi^X, (P^X g_1)^2) \\ &= \pi(w) - \pi(w-) P(w_{h-1}, w_h)^2 \\ &= \pi(w)[1 - P(w_{h-1}, w_h)], \end{aligned}$$

car $\pi(w) = \pi(w-)P(w_{h-1}, w_h)$. Remarquons ensuite que l'on a $g_1 g_2 = g_1$ et $(P^X g_1)(P^X g_2) = (P^X g_1)^2$, et donc $\Sigma_{12} = \Sigma_{21} = \Sigma_{11}$. Enfin, comme $g_2^2 = g_2$, on obtient, avec $\pi(w_{h-1}w_h) = \pi(w_{h-1})P(w_{h-1}, w_h)$ (cf. la définition (6.1)) :

$$\begin{aligned} \Sigma_{22} &= (\pi^X, P^X(g_2)) - (\pi^X, (P^X g_2)^2) \\ &= \pi(w_{h-1}w_h) - \pi(w_{h-1}) P(w_{h-1}, w_h)^2 \\ &= \pi(w_{h-1}w_h)[1 - P(w_{h-1}, w_h)]. \end{aligned}$$

Il vient donc

$$\Sigma = [1 - P(w_{h-1}, w_h)] \begin{pmatrix} \pi(w) & \pi(w) \\ \pi(w) & \pi(w_{h-1}w_h) \end{pmatrix}.$$

On déduit du corollaire 1.5.11 que la suite $(N_{w-}/N_{w_{h-1}}, N \geq h)$ converge p.s. vers $\pi(w-)/\pi(w_{h-1})$, qui est bien défini car $\pi(w_{h-1}) > 0$. Cela implique, grâce au théorème de Slutsky, la convergence en loi du vecteur $\big(H_N, N_{w-}/N_{w_{h-1}}, N \geq h\big)$ vers $(G, \pi(w-)/\pi(w_{h-1}))$. On pose

$$\zeta_N = \left(1, -\frac{N_{w-}}{N_{w_{h-1}}}\right) H_N = \frac{1}{\sqrt{N}} \left(N_w - \frac{N_{w-} N_{w_{h-1}w_h}}{N_{w_{h-1}}}\right).$$

L'application $f : ((h_1, h_2), x) \to h_1 - x h_2$ est une application continue de $\mathbb{R}^2 \times \mathbb{R}$ dans $\mathbb{R}$. Donc $\Big(\zeta_N = f(H_N, N_{w-}/N_{w_{h-1}}), N \geq h\Big)$ converge en loi vers $f(G, \pi(w-)/\pi(w_{h-1})) = (1, -\pi(w-)/\pi(w_{h-1}))G$ de loi gaussienne centrée et de variance

$$\begin{aligned} \sigma^2 &= \big(1, -\pi(w-)/\pi(w_{h-1})\big) \Sigma \begin{pmatrix} 1 \\ -\pi(w-)/\pi(w_{h-1}) \end{pmatrix} \\ &= \big(1, -\pi(w-)/\pi(w_{h-1})\big) \begin{pmatrix} \pi(w)\big[1 - \frac{\pi(w-)}{\pi(w_{h-1})}\big] \\ 0 \end{pmatrix} [1 - P(w_{h-1}, w_h)] \\ &= \pi(w)\Big[1 - \frac{\pi(w-)}{\pi(w_{h-1})}\Big][1 - P(w_{h-1}, w_h)], \end{aligned} \tag{6.4}$$

où l'on a utilisé $\pi(w) = \dfrac{\pi(w-)\pi(w_{h-1}w_h)}{\pi(w_{h-1})}$ pour la deuxième égalité. □

6.2 Une autre approche asymptotique

On rappelle que $w = w_1 \cdots w_h$ est un mot de longueur $h \geq 3$. Si l'on considère (6.2), il est naturel de comparer N_w/N et sa limite (inconnue) $\pi(w)$. Dans le paragraphe précédent, la probabilité $\pi(w)$ a été estimée par l'estimateur convergent $\frac{1}{N}\,\frac{N_w - N_{w_{h-1}w_h}}{N_{w_{h-1}}}$ (voir la remarque 6.1.3, point 3). Il apparaît en fait plus naturel de considérer l'estimateur du maximum de vraisemblance (EMV) de $\pi(w)$.

On commence par le lemme préliminaire suivant, dont on pourra survoler la démonstration qui est reportée à la fin de ce paragraphe.

Lemme 6.2.1. *La suite* $\big(\big(\sqrt{N}\big[\frac{N_{yy'}}{N_y} - P(y,y')\big],\ \ y,y' \in E\big), N \geq 2\big)$ *converge en loi vers un vecteur gaussien centré dont la matrice de covariance* $\Sigma = (\Sigma_{xx',yy'},\ x,x',y,y' \in E)$ *est définie par*

$$\Sigma_{xx',yy'} = \frac{1}{\pi(x)}\, P(x,x')[\mathbf{1}_{\{x'=y'\}} - P(y,y')]\mathbf{1}_{\{x=y\}}.$$

De plus $\hat{P}_N = (\hat{P}_N(y,y') = N_{yy'}/N_y,\ y,y' \in E)$ *est un estimateur convergent de* P*, asymptotiquement normal de même variance asymptotique que l'estimateur du maximum de vraisemblance de* P*.*

Par abus de langage, on dira que $\hat{P}_N$ est l'EMV de P.

D'un point de vue pratique, signalons que comme $N_y = \sum_{z\in E} N_{yz} + \mathbf{1}_{\{y_N=y\}} = \sum_{z\in E} N_{zy} + \mathbf{1}_{\{y_1=y\}}$, le calcul de $\hat{P}_N$ ne nécessite que la connaissance des nombres d'occurrences des mots de deux lettres et la valeur de la première base (i.e. y_1) de la séquence considérée. En fait, l'ensemble des nombres d'occurrences observées des mots de deux lettres et la valeur de la première lettre forme un ensemble (on parle de statistique) qui contient toute l'information suffisante à l'estimation des paramètres du modèle, c'est-à-dire de la matrice de transition. Plus précisément, on peut montrer que la loi de $(Y_1,\ldots,Y_N)$, sachant la statistique Y_1 et $(N_{yy'}, y,y' \in E)$, ne dépend pas du paramètre P. On dit que la statistique est exhaustive. Cela implique en particulier que, dans le cadre du modèle considéré, il est cohérent d'écrire tous les estimateurs et tous les tests à l'aide de cette statistique exhaustive. De plus les estimateurs construits à partir des statistiques exhaustives possèdent en général de bonnes propriétés.

Pour tenir compte de ces remarques, on observe que pour un mot $v = v_1 \cdots v_k$ de longueur $k \geq 2$, la suite d'estimateurs construits à partir de la statistique exhaustive, $(\hat{\pi}_N(v), N \geq h)$, où

$$\hat{\pi}_N(v) = \frac{N_{v_1}}{N} \prod_{l=1}^{k-1} \frac{N_{v_l v_{l+1}}}{N_{v_l}} = \frac{1}{N}\, \frac{N_{v_1 v_2} \cdots N_{v_{k-1} v_k}}{N_{v_2} \cdots N_{v_{k-1}}}, \tag{6.5}$$

converge p.s. vers $\pi(v)$ défini par (6.1). On explique dans la remarque 6.2.6 que $\hat{\pi}_N(v)$ est en fait l'EMV de $\pi(v)$.

On considère maintenant l'écart entre N_w/N et $\hat{\pi}_N(w)$ en posant

$$\zeta'_N = \frac{1}{\sqrt{N}}(N_w - N\hat{\pi}_N(w)).$$

Avant de donner le théorème de convergence concernant ζ'_N, on introduit quelques notations liées au fait que les mots w peuvent se chevaucher.

Pour $d \in \{1, \ldots, h-1\}$, on note $\delta(w\,;d) = 1$ si $w = w_1 \cdots w_d w_1 \cdots w_{h-d}$ (i.e. si le mot w peut apparaître simultanément en position i et $i+d$), et $\delta(w\,;d) = 0$ sinon. Si $\delta(w\,;d) = 1$, alors on considérera le mot de longueur $h+d$: $w^{(d)}w = w_1 \cdots w_d w_1 \cdots w_h$. (Si on considère le mot $w = aba$, alors on a $\delta(w\,;1) = 0$, $\delta(w\,;2) = 1$ et $w^{(2)}w = ababa$.) Enfin, on note $n_v(w')$ le nombre d'occurrences du mot v dans le mot w'.

On rappelle que $w- = w_1 \cdots w_{h-1}$ désigne le mot w tronqué de sa dernière lettre. On admet le théorème suivant (voir [8 et 11], où ce théorème est démontré dans un cadre plus général).

Théorème 6.2.2. *Avec les notations qui précèdent, la suite* $(\zeta'_N, N \geq h)$ *converge en loi vers* G *de loi gaussienne* $\mathcal{N}(0, \sigma'^2)$*, où*

$$\sigma'^2 = \pi(w) + 2\sum_{d=1}^{h-2} \delta(w\,;d)\pi(w^{(d)}w) + \pi(w)^2 \left(\sum_{y \in E} \frac{n_y(w-)^2}{\pi(y)} - \sum_{y,z \in E} \frac{n_{yz}(w)^2}{\pi(yz)} - \frac{2n_{w_1}(w-) - 1}{\pi(w_1)} \right).$$

Le deuxième terme dans la définition de σ'^2 provient du fait que les mots w peuvent se chevaucher. Le chevauchement possible des mots rend délicates les démonstrations des théorèmes asymptotiques.

D'après les commentaires qui suivent (6.5), l'estimateur suivant est un estimateur convergent de σ'^2 :

$$\hat{\sigma}'^2_N = \hat{\pi}_N(w) + 2\sum_{d=1}^{h-2} \delta(w\,;d)\hat{\pi}_N(w^{(d)}w) + \hat{\pi}_N(w)^2 \left(\sum_{y \in E} \frac{n_y(w-)^2}{\hat{\pi}_N(y)} - \sum_{y,z \in E} \frac{n_{yz}(w)^2}{\hat{\pi}_N(yz)} - \frac{2n_{w_1}(w-) - 1}{\hat{\pi}_N(w_1)} \right).$$

On pose $\hat{\sigma}'_N = \sqrt{\hat{\sigma}'^2_N}$. On déduit du théorème de Slutsky et du théorème 6.2.2 le corollaire suivant.

Corollaire 6.2.3. *La suite* $Z' = (Z'_N = \zeta'_N/\hat{\sigma}'_N, N \geq h)$ *converge en loi vers* G *de loi gaussienne centrée réduite.*

On peut alors reproduire le raisonnement qui suit le corollaire 6.1.2, utilisant les p-valeurs construites à l'aide de la statistique Z'_N, pour exhiber les mots w exceptionnellement fréquents ou rares d'une séquence observée. Ici le test compare le nombre d'occurrences d'un mot w avec le nombre d'occurrences des mots de une et deux lettres qui le composent, contrairement au test du paragraphe 6.1 qui compare le nombre d'occurrences du mot w avec le nombre d'occurrences de $w-$ et du mot formé de ses deux lettres finales $w_{h-1}w_h$. Ces deux approches sont différentes si $h > 3$. Enfin, pour des mots w de longueur $h = 3$, l'exercice qui suit permet de se convaincre que l'approche de ce paragraphe (théorème 6.2.2) et celle du paragraphe précédent (théorème 6.1.1) coïncident. En revanche les corollaires 6.2.3 et 6.1.2 proposent des estimations de σ^2 différentes.

Exercice 6.2.4. On considère les mots de trois lettres : on suppose $h = 3$. Montrer que $\hat{\pi}_N(w) = N_{w-}N_{w_{h-1}w_h}/N_{w_{h-1}}$. En particulier, on a $\zeta_N = \zeta'_N$ (voir (6.3) pour la définition de ζ_N). Remarquer que si $\delta(w\,;1) = 1$, alors il existe $a \in E$ tel que $w = aaa$. Montrer, en distinguant suivant les cas $\delta(w\,;1) = 1$ et $\delta(w\,;1) = 0$, que σ'^2 défini dans le théorème 6.2.2 peut se récrire de la manière suivante :

$$\sigma'^2 = \pi(w) - \pi(w)^2 \left(\frac{1}{\pi(w_1w_2)} + \frac{1}{\pi(w_2w_3)} - \frac{1}{\pi(w_2)} \right).$$

Vérifier que la variance σ^2 définie par (6.4) est égale à σ'^2.

Ainsi pour les mots de longueur 3, le théorème 6.2.2 et le théorème 6.1.1 sont identiques. En revanche les théorèmes diffèrent si l'on considère des mots de longueur $h > 3$. ♦

Exemple 6.2.5. Les figures 6.1 et 6.2 présentent les histogrammes des variables Z_N (définies au paragraphe précédent) et Z'_N calculées pour tous les mots de longueur 6 et 8, correspondant à la simulation d'une chaîne de Markov à valeurs dans $E = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ avec $N = 4\,639\,221$ (N correspond à la longueur de la séquence d'ADN de E. coli).

La matrice de transition utilisée pour les simulations est proche de celle estimée pour E. coli (voir (6.11)) :

$$P = \begin{pmatrix} 0.30 & 0.22 & 0.21 & 0.27 \\ 0.27 & 0.22 & 0.31 & 0.20 \\ 0.22 & 0.32 & 0.24 & 0.22 \\ 0.18 & 0.23 & 0.30 & 0.29 \end{pmatrix}.$$

Il existe $4^6 = 4\,096$ mots distincts de longueur 6 et $4^8 = 65\,536$ mots distincts de longueur 8. Bien que les variables Z_N (ainsi que Z'_N) ne soient pas indépendantes pour tous les mots, on observe une bonne adéquation entre les histogrammes et la densité de la loi gaussienne centrée réduite.

La figure 6.3 (resp. 6.4) présente pour la même simulation les points de coordonnées (Z_N, Z'_N) pour tous les mots de longueur 6 (resp. 8). On remarque

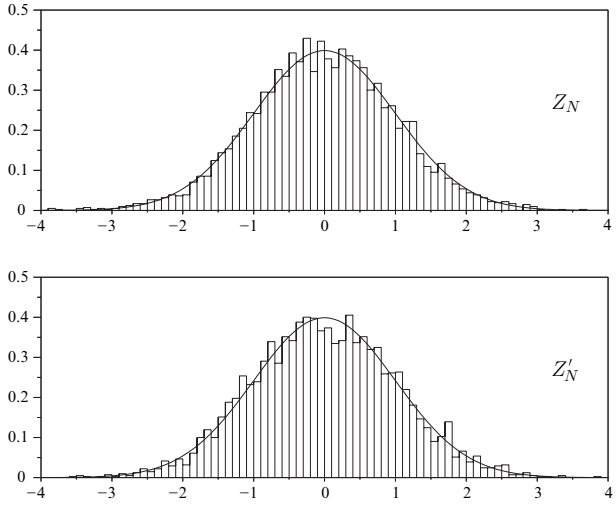

Fig. 6.1. Histogrammes de Z_N et Z'_N pour tous les mots de longueur 6 observés sur une simulation d'un ADN de même longueur que celui de E. coli, comparé avec la densité de la loi $\mathcal{N}(0,1)$

que le nuage de points est positionné autour de la diagonale, et que les valeurs typiques varient dans $[-5,5]$. Les p-valeurs approchées pour les mots de 6 lettres, calculées à partir du Tableau 6.1 page 192, sont comprises entre 0.00007 et 0.9998 pour Z_N et entre 0.0002 et 0.99995 pour Z'_N. $\Diamond$

Démonstration du lemme 6.2.1. On considère la fonction vectorielle sur E^2 : $h = (h_{yy'} = \mathbf{1}_{\{(y,y')\}}, y, y' \in E)$. On a

$$Ph_{yy'}(x) = \sum_{z\in E} P(x,z)\mathbf{1}_{\{(y,y')\}}(x,z) = P(y,y')\mathbf{1}_{\{y\}}(x).$$

Remarquons que $\sum_{k=2}^{N} h_{yy'}(X_{k-1}, X_k) = N_{yy'}$ et

$$\sum_{k=2}^{N} Ph_{yy'}(X_{k-1}) = N_y P(y,y') - \mathbf{1}_{\{Y_N=y\}}P(y,y').$$

On déduit du corollaire 1.6.5 la convergence en loi suivante :

$$\Big(\frac{1}{\sqrt{N}} \big[N_{yy'} - N_y P(y,y') + \mathbf{1}_{\{Y_N=y\}}P(y,y') \big], y, y' \in E \Big) \xrightarrow[N\to\infty]{\textbf{Loi}} \mathcal{N}(0,\Sigma'),$$

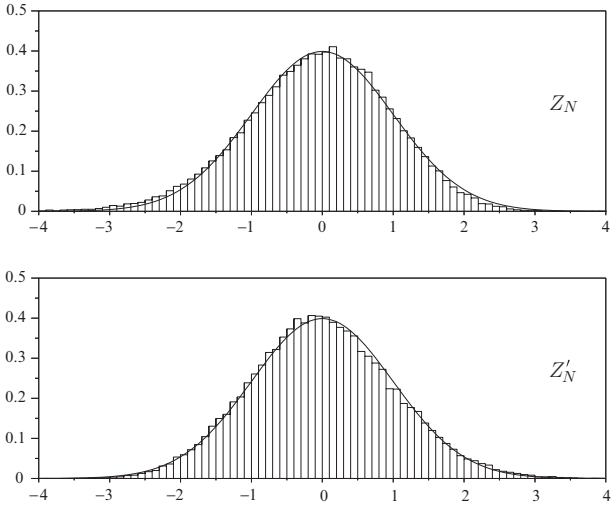

Fig. 6.2. Histogrammes de Z_N et Z'_N pour les mots de longueur 8 observés sur une simulation d'un ADN de même longueur que celui de E. coli, comparé avec la densité de la loi $\mathcal{N}(0,1)$

où la matrice $\Sigma' = (\Sigma'_{xx',yy'}, x, x', y, y' \in E)$ est définie par

$$\begin{aligned}\Sigma'_{xx',yy'} &= (\pi, P(h_{xx'}h_{yy'})) - (\pi, (Ph_{xx'})(Ph_{yy'})) \\ &= \pi(x)P(x,x')\mathbf{1}_{\{xx'=yy'\}} - \pi(x)P(x,x')P(y,y')\mathbf{1}_{\{x=y\}} \\ &= \mathbf{1}_{\{x=y\}}\pi(x)P(x,x')[\mathbf{1}_{\{x'=y'\}} - P(y,y')].\end{aligned}$$

Comme les suites $(\mathbf{1}_{\{Y_N=y\}}P(y,y')/\sqrt{N}, N \geq 2)$ convergent p.s. vers 0 pour tout $y \in E$, on en déduit que

$$\left(\frac{N_y}{N}\sqrt{N}\left[\frac{N_{yy'}}{N_y} - P(y,y')\right], y, y' \in E\right) \xrightarrow[N\to\infty]{\textbf{Loi}} \mathcal{N}(0,\Sigma').$$

Le théorème ergodique implique la convergence p.s. des suites $(N/N_y, N \geq 1)$ vers $1/\pi(y)$ pour $y \in E$. On déduit alors du théorème de Slutsky que

$$\left(\sqrt{N}\left[\frac{N_{yy'}}{N_y} - P(y,y')\right], y, y' \in E\right) \xrightarrow[N\to\infty]{\textbf{Loi}} \mathcal{N}(0,\Sigma),$$

où $\Sigma = (\Sigma_{xx',yy'} = \Sigma'_{xx',yy'}/[\pi(x)\pi(y)], x, x', y, y' \in E)$. On obtient ainsi la première partie du lemme.

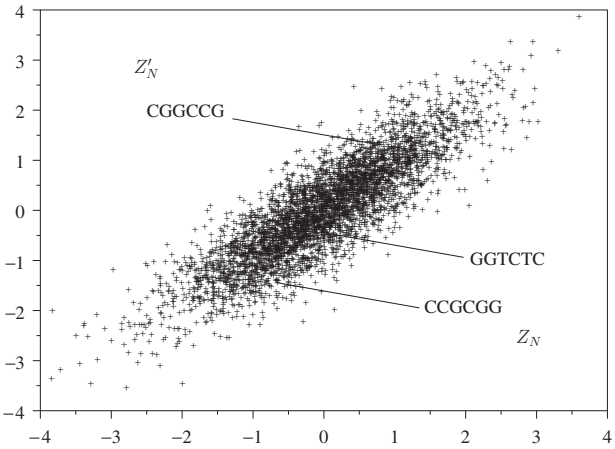

Fig. 6.3. (Z_N, Z'_N) pour tous les mots de longueur 6 observés sur une simulation d'un ADN de même longueur que celui de E. coli

On recherche maintenant l'EMV, $\tilde{P}_N$, de la matrice de transition P. La vraisemblance associée à la chaîne de Markov $(Y_n, n \in \{1, \dots, N\})$ est

$$\begin{aligned} p_N(P\,; y_1, \dots, y_n) &= \mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) \\ &= \mathbb{P}(Y_1 = y_1) P(y_1, y_2) \cdots P(y_{N-1}, y_N) \\ &= \mathbb{P}(Y_1 = y_1) \prod_{y,y' \in E} P(y, y')^{N_{yy'}}, \end{aligned}$$

où $N_{yy'}$ est le nombre d'occurrences du mot yy'. On en déduit la log-vraisemblance :

$$\begin{aligned} L_N(P\,; y_1, \dots, y_N) &= \log p_N(P\,; y_1, \dots, y_n) \\ &= \log(\mathbb{P}(Y_1 = y_1)) + \sum_{y,y' \in E} N_{yy'} \log(P(y, y')). \end{aligned}$$

L'EMV de P est la matrice $\tilde{P}_N = (\tilde{P}_N(y, y'), y, y' \in E)$ qui maximise la log-vraisemblance et telle que $(\tilde{P}_N(y, y'), y' \in E)$ est une probabilité pour tout $y \in E$. Comme ces contraintes sont séparées pour $y \in E$, on en déduit que pour tout $y \in E$, on recherche la probabilité $(\tilde{P}_N(y, y'), y' \in E)$ qui maximise $\sum_{y' \in E} N_{yy'} \log(P(y, y'))$, et donc qui maximise $\sum_{y' \in E} p(y') \log(P(y, y'))$, où $(p(y') = N_{yy'} / \sum_{z \in E} N_{yz}, y' \in E)$ est une probabilité sur E. On déduit du lemme 5.2.8 que, sous la contrainte que $(P(y, y'), y' \in E)$ soit une probabilité, la quantité $\sum_{y' \in E} p(y') \log(P(y, y'))$ est maximale pour $P(y, y') = p(y')$ pour tout $y' \in E$. Ainsi, pour $y, y' \in E$, l'EMV de $P(y, y')$ est

$$\tilde{P}_N(y, y') = \frac{N_{yy'}}{\sum_{z \in E} N_{yz}}.$$

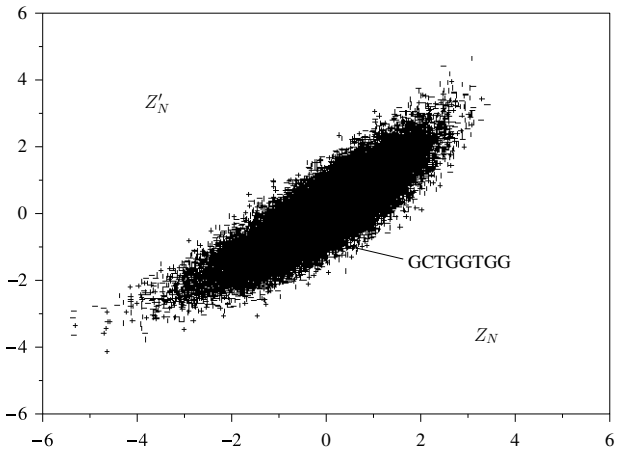

Fig. 6.4. (Z_N, Z'_N) pour tous les mots de longueur 8 observés sur une simulation d'un ADN de même longueur que celui de E. coli

Pour vérifier que $\tilde{P}_N$ et $\hat{P}_N$ ont même variance asymptotique, il suffit de vérifier que p.s. pour tous $y, y' \in E$,

$$\lim_{N\to\infty} N^{1/2}[\tilde{P}_N(y,y') - \hat{P}_N(y,y')] = 0. \tag{6.6}$$

Comme $\sum_{z\in E} N_{yz} = N_y - \mathbf{1}_{\{y_N=y\}}$, on a

$$\tilde{P}_N(y,y') - \hat{P}_N(y,y') = \frac{N_{yy'}}{N_y \sum_{z\in E} N_{yz}} \mathbf{1}_{\{y_N=y\}}.$$

On déduit du corollaire 1.5.11, que p.s. $\lim_{N\to\infty} N \dfrac{N_{yy'}}{N_y \sum_{z\in E} N_{yz}} = \pi(yy')/\pi(y)^2$. Cette limite étant fini, cela implique (6.6) et termine la démonstration du lemme. □

Remarque 6.2.6. L'estimateur $\hat{\pi}_N(v)$ de $\pi(v)$, défini par (6.5), est un estimateur convergent. Nous allons vérifier qu'il est également asymptotiquement normal de même variance que l'EMV de $\pi(v)$.

Dans une première étape, on vérifie que la probabilité invariante π peut s'écrire comme une fonction régulière de P. On rappelle le théorème de Perron-Frobenius (voir par exemple [12]).

Théorème 6.2.7. *Soit $P' = (P'(y,y'), (y,y') \in E^2)$ une matrice dont tous les coefficients sont strictement positifs. Alors elle possède une valeur propre réelle positive simple, $\lambda_{P'}$, strictement plus grande que le module de toutes les autres valeurs propres de P'. Le vecteur propre à gauche, $\pi_{P'}$, associé à la*

valeur propre $\lambda_{P'}$ peut être normalisé de telle sorte que $\pi_{P'} = (\pi_{P'}(y) > 0, y \in E)$ soit une probabilité sur E. Si de plus P' est une matrice stochastique, alors $\lambda_{P'} = 1$.

Ainsi si P' est une matrice stochastique dont les coefficients sont strictement positifs, alors $\pi_{P'}$ est la probabilité invariante de la chaîne de Markov de matrice de transition P'.

La matrice P est une matrice stochastique dont tous les coefficients sont strictement positifs, et bien sûr $\pi_P = \pi$. Comme λ_P est racine simple, il existe un voisinage ouvert de P dans $\mathbb{R}^{E\times E}$, O, et une fonction φ définie sur O de classe au moins $\mathcal{C}^1$ tels que si $P' \in O$, alors P' est une matrice dont les coefficients sont strictement positifs et $\varphi(P') = \pi_{P'}$.

Dans une deuxième étape on exhibe un estimateur proche de l'EMV de (P, π). On reprend les notations de la démonstration du lemme 6.2.1. Pour N assez grand l'EMV, $\tilde{P}_N$, de P appartient à O. Par convention (voir la définition 5.2.2), l'EMV de π est l'image par φ de l'EMV de P, c'est-à-dire $\tilde{\pi}_N = \varphi(\tilde{P}_N)$, la probabilité invariante de $\tilde{P}_N$. Comme φ est de classe $\mathcal{C}^1$, la proposition A.3.17 implique que $\tilde{\pi}_N$ est un estimateur asymptotiquement normal de π, et plus généralement que l'EMV de (P, π), $(\tilde{P}_N, \tilde{\pi}_N)$, est un estimateur asymptotiquement normal.

En fait il est naturel de choisir $\hat{\pi}_N = (\hat{\pi}_N(y) = N_y/N, y \in E)$ comme estimateur de la probabilité invariante π. Cet estimateur est convergent grâce au théorème ergodique. Il est facile de vérifier que $\hat{\pi}_N$ est la probabilité invariante de la matrice stochastique $\check{P}_N$ définie pour $y, y' \in E$ par

$$\check{P}_N(y, y') = \frac{N_{yy'} + \mathbf{1}_{\{y_N=y, y_1=y'\}}}{N_y}.$$

Nous vérifions ensuite que $\check{P}_N$ est proche de $\tilde{P}_N$. On a

$$\tilde{P}_N(y, y') - \check{P}_N(y, y') = \frac{N_{yy'}}{N_y \sum_{z\in E} N_{yz}} \mathbf{1}_{\{y_N=y\}} - \frac{1}{N_y} \mathbf{1}_{\{y_N=y, y_1=y'\}}.$$

Ainsi l'égalité (6.6) est satisfaite avec $\check{P}_N$ au lieu de $\hat{P}_N$. Pour N assez grand, on a $\check{P}_N \in O$, et $\hat{\pi}_N = \varphi(\check{P}_N)$. Comme φ est au moins de classe $\mathcal{C}^1$ sur un voisinage de P, on en déduit alors que p.s.

$$\lim_{N\to\infty} N^{1/2}[\tilde{\pi}_N(y) - \hat{\pi}_N(y)] = 0.$$

En particulier ceci assure que $\hat{\pi}_N$ est asymptotiquement normal de même variance asymptotique que $\tilde{\pi}_N$. En fait, avec (6.6), on obtient que $(\hat{P}_N, \hat{\pi}_N)$ est un estimateur asymptotiquement normal de (P, π), de même variance asymptotique que l'EMV.

On en déduit que $\hat{\pi}_N(v) = \hat{\pi}(v_1) \prod_{l=1}^{k-1} \hat{P}(v_l, v_{l+1})$ est un estimateur asymptotiquement normal de $\pi(v) = \pi(v_1) \prod_{l=1}^{k-1} P(v_l, v_{l+1})$ de même variance asymptotique que l'EMV. De fait, on dira que $\hat{\pi}_N(v)$ est l'EMV de $\pi(v)$. $\Diamond$

6.3 Une troisième approche asymptotique

Comme nous l'avons souligné dans la remarque 6.1.3, point 2, les résultats du type TCL, comme ceux énoncés dans le corollaire 6.1.2 et le corollaire 6.2.3 ne donnent pas de bonnes approximations de la p-valeur pour des mots ayant une faible probabilité d'apparition, ce qui est le cas par exemple si la probabilité, $\pi(w)$, d'observer le mot w est de l'ordre de $1/N$, où N est la longueur de la séquence observée. Intuitivement, le nombre d'occurrences du mot w sera alors de l'ordre de $N\pi(w) = O(1)$. On n'est pas dans le régime de la loi forte des grands nombres (et donc pas dans le cadre du TCL), mais plutôt dans un régime de type événements rares ou « loi des petits nombres » (voir [1] où de nombreux exemples sont traités concernant la « loi des petits nombres »).

Avant de donner les résultats concernant l'analyse du nombre des occurrences des mots dans l'optique de la «loi des petits nombres», nous présentons d'abord quelques résultats élémentaires sur cette loi.

6.3.1 « Loi des petits nombres » ou loi de Poisson

L'exemple élémentaire suivant rappelle comment la loi de Poisson apparaît naturellement comme « loi des petits nombres ».

Exemple 6.3.1. Soit $X_1^n = (X_k^n, k \in \{1, \dots, n\})$ des variables aléatoires de Bernoulli de même paramètre p_n, indépendantes. On note $S_n = \sum_{k=1}^n X_k^n$ le nombre d'occurrences de 1 dans la suite X_1^n. La loi de S_n est la loi binomiale de paramètre (n, p_n) : pour $k \in \{0, \dots, n\}$,

$$\mathbb{P}(S_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{1}{k!}\, \mathrm{e}^{(n-k)\log(1-p_n)} \prod_{i=1}^{k} (n-i+1)p_n.$$

En particulier, si $\lim_{n\to\infty} np_n = \theta \in]0, \infty[$, on obtient que pour tout $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{P}(S_n = k) = \frac{1}{k!}\, \mathrm{e}^{-\theta}\, \theta^k.$$

La suite $(S_n, n \geq 1)$ converge donc en loi vers une variable de loi de Poisson de paramètre θ. Ce résultat est un exemple de la « loi des petits nombres » : quand la probabilité d'un événement, p_n, est de l'ordre de $1/n$, et que l'on dispose de n observations, le nombre d'occurrences de l'événement suit asymptotiquement une loi de Poisson. ◊

Dans la suite de ce paragraphe nous montrons comment ce résultat peut s'étendre à une suite de variables aléatoires de Bernoulli, $(X_n, n \geq 1)$, indépendantes mais pas de même loi. Pour cela, on désire comparer, au sens de la norme en variation (voir l'appendice D), la loi de $S_n = \sum_{k=1}^n X_k$ et la loi de Poisson de paramètre $\theta_n = \sum_{k=1}^n p_k$, où $p_k = \mathbb{P}(X_k = 1)$ est le paramètre de la loi de Bernoulli de X_k.

En utilisant les fonctions génératrices, il est immédiat de vérifier que la loi de $U_n = \sum_{k=1}^n V_k$, où les variables V_k sont indépendantes de loi de Poisson de paramètres respectifs p_k, est la loi de Poisson de paramètre $\theta_n = \sum_{k=1}^n p_k$.

Nous majorons la distance, pour la norme en variation, entre la loi de Bernoulli de paramètre p_n et la loi de Poisson de paramètre p_n. Pour cela, on note de manière générale μ_T, la loi d'une variable aléatoire à valeurs entières $T : \mu_T(k) = \mathbb{P}(T = k)$ pour $k \in \mathbb{N}$. On a

$$\begin{aligned}\|\mu_{X_n} - \mu_{V_n}\| &= \frac{1}{2}\left[\left|1 - p_n - \mathrm{e}^{-p_n}\right| + p_n(1 - \mathrm{e}^{-p_n}) + \sum_{i\geq 2}\frac{1}{i!}\,p_n^i\,\mathrm{e}^{-p_n}\right] \\ &= \frac{1}{2}\left[\mathrm{e}^{-p_n} + p_n - 1 + p_n(1 - \mathrm{e}^{-p_n}) + 1 - \mathrm{e}^{-p_n} - p_n\,\mathrm{e}^{-p_n}\right] \\ &= p_n(1 - \mathrm{e}^{-p_n}) \\ &\leq p_n^2,\end{aligned}$$

où l'on a utilisé que $\mathrm{e}^{-x} - 1 + x \geq 0$ pour $x \geq 0$, dans l'inégalité.

Le lemme suivant permet de majorer la distance, pour la norme en variation, entre des lois de sommes de variables aléatoires indépendantes.

Lemme 6.3.2. *Soit $(X'_k, 1 \leq k \leq n)$ et $(V'_k, 1 \leq k \leq n)$ deux suites de variables aléatoires indépendantes à valeurs dans $\mathbb{N}$. On a*

$$\left\|\mu_{\sum_{k=1}^n X'_k} - \mu_{\sum_{k=1}^n V'_k}\right\| \leq \sum_{k=1}^n \left\|\mu_{X'_k} - \mu_{V'_k}\right\|.$$

Démonstration. Si on établit le résultat pour $n = 2$, alors un raisonnement par récurrence évident permet d'obtenir le résultat pour n quelconque. On a

$$\begin{aligned}&\left\|\mu_{X'_1+X'_2} - \mu_{V'_1+V'_2}\right\| \\ &\quad= \frac{1}{2}\sum_{i\in\mathbb{N}}\left|\mathbb{P}(X'_1 + X'_2 = i) - \mathbb{P}(V'_1 + V'_2 = i)\right| \\ &\quad= \frac{1}{2}\sum_{i\in\mathbb{N}}\left|\sum_{l=0}^{i}[\mathbb{P}(X'_1 = l, X'_2 = i - l) - \mathbb{P}(V'_1 = l, V'_2 = i - l)]\right| \\ &\quad= \frac{1}{2}\sum_{i\in\mathbb{N}}\left|\sum_{l=0}^{i}[\mathbb{P}(X'_1 = l)\mathbb{P}(X'_2 = i - l) - \mathbb{P}(V'_1 = l)\mathbb{P}(V'_2 = i - l)]\right| \\ &\quad\leq \frac{1}{2}\sum_{i\in\mathbb{N}}\sum_{l=0}^{i}\left|\mathbb{P}(X'_1 = l)\mathbb{P}(X'_2 = i - l) - \mathbb{P}(V'_1 = l)\mathbb{P}(V'_2 = i - l)\right| \\ &\quad\leq \frac{1}{2}\sum_{j,l\in\mathbb{N}}\left|\mathbb{P}(X'_1 = l)\mathbb{P}(X'_2 = j) - \mathbb{P}(V'_1 = l)\mathbb{P}(V'_2 = j)\right|,\end{aligned}$$

où on a posé $j = i - l$ pour la dernière inégalité. Il vient

$$\begin{aligned}\left\|\mu_{X_1'+X_2'} - \mu_{V_1'+V_2'}\right\| &\leq \frac{1}{2}\sum_{j,l\in\mathbb{N}} |\mathbb{P}(X_1' = l) - \mathbb{P}(V_1' = l)|\,\mathbb{P}(X_2' = j) \\ &\qquad + \frac{1}{2}\sum_{j,l\in\mathbb{N}} |\mathbb{P}(X_2' = j) - \mathbb{P}(V_2' = j)|\,\mathbb{P}(V_1' = l) \\ &= \left\|\mu_{X_1'} - \mu_{V_1'}\right\| + \left\|\mu_{X_2'} - \mu_{V_2'}\right\|.\end{aligned}$$

Ceci termine la démonstration du lemme. □

On en déduit que

$$\|\mu_{S_n} - \mu_{U_n}\| \leq \sum_{k=1} \|\mu_{X_k} - \mu_{V_k}\| \leq \sum_{k=1}^{n} p_k^2. \tag{6.7}$$

En particulier, dans le cas où les paramètres p_k sont tous égaux à p, on obtient $\left\|\mu_{\sum_{k=1}^n X_k} - \mu_{\sum_{k=1}^n V_k}\right\| \leq np^2$. Pour $p = \theta/n$, la distance pour la norme en variation entre la loi binomiale de paramètre $(n, \theta/n)$ et la loi de Poisson de paramètre θ est majorée par θ^2/n. Elle tend vers 0 quand $n \to \infty$. On retrouve ainsi le résultat de l'exemple 6.3.1 ci-dessus, pour $p_n = \theta/n$.

De nombreux résultats plus précis que (6.7) sur l'approximation de la loi de la somme de variables de Bernoulli indépendantes par une loi de Poisson existent, ainsi que des résultats dans le même esprit concernant la somme de variables de Bernoulli dépendantes, voir par exemple [2].

Enfin, l'exercice qui suit permet de retrouver complètement le résultat élémentaire de l'exemple 6.3.1.

Exercice 6.3.3. Soit U et U' des variables aléatoires de Poisson de paramètres respectifs θ et θ'. Vérifier que si $\theta \geq \theta' > 0$, alors on a

$$|\,\mathrm{e}^{-\theta}\,\theta^k - \mathrm{e}^{-\theta'}\,\theta'^k| \leq \mathrm{e}^{-\theta}(\theta^k - \theta'^k) + \theta'^k(\mathrm{e}^{-\theta'} - \mathrm{e}^{-\theta}).$$

En déduire que $\|\mu_U - \mu_{U'}\| \leq 1 - \mathrm{e}^{-|\theta-\theta'|}$. Montrer que la distance pour la norme en variation entre la loi binomiale de paramètre (n, p_n) et la loi de Poisson de paramètre θ est majorée par $np_n^2 + 1 - \mathrm{e}^{-|\theta-np_n|}$. Retrouver ainsi le résultat de l'exemple 6.3.1. ♦

6.3.2 « Loi des petits nombres » pour le nombre d'occurrences

Nous reprenons les hypothèses et notations du paragraphe 6.1 : la séquence $y_1, \ldots, y_N$ est la réalisation des N premiers termes d'une chaîne de Markov sur E (fini non réduit à un singleton) irréductible $(Y_n, n \geq 1)$ de matrice de transition P à coefficients strictement positifs et de probabilité invariante π.

Soit $w = w_1 \cdots w_h$ un mot de longueur $h \geq 3$. On note $V_i = 1$ si le mot w commence en position i de la séquence et 0 sinon :

$$V_i = \mathbf{1}_{\{Y_i^{i+h-1} = w\}}.$$

Le nombre d'occurrences de w peut alors s'écrire comme $N_w = \sum_{i=1}^{N-h+1} V_i$. En régime stationnaire (i.e. si la loi de Y_1 est la probabilité invariante π), la loi de V_i est la loi de Bernoulli de paramètre $\pi(w)$. Donc N_w est la somme de N variables (dépendantes) de loi de Bernoulli de paramètre $\pi(w)$.

Si $\pi(w)$ est de l'ordre de $1/N$, alors d'après le paragraphe précédent, la loi du nombre d'occurrences du mot w est à comparer avec une loi de Poisson, même si on s'attend à des phénomènes plus complexes dus à la dépendance des variables $(V_i, i \in \{1, \dots N\})$ entre elles.

Si le mot w ne peut pas se chevaucher lui-même (i.e., avec les notations précédant le théorème 6.1.1, si $\delta(w\,; d) = 0$ pour $d \in \{1, \dots, h-1\}$), alors si $N\pi(w) = O(1)$, on peut montrer que la loi de $N(w)$ est proche de la loi de Poisson de paramètre $N\pi(w)$. En revanche, si le mot w peut se chevaucher avec lui-même, alors il faut étudier le nombre d'occurrences de groupes de mots w se chevauchant. On note $\tilde{V}_i = 1$ si un mot w commence en position i et si aucun mot commençant avant la position i ne le chevauche, et $\tilde{V}_i = 0$ sinon :

$$\tilde{V}_i = V_i \prod_{k=1}^{\min(h,i)-1} (1 - V_{i-k}).$$

On dit qu'un train de mots w débute en position i si $\tilde{V}_i = 1$. Le nombre de trains observés est donc

$$\tilde{N}_w = \sum_{i=1}^{N-h+1} \tilde{V}_i.$$

(Pour la séquence *aabaaa* et le mot $w = aa$, on a $V_1 = V_4 = V_5 = 1$, mais $\tilde{V}_1 = \tilde{V}_4 = 1$ et $\tilde{V}_5 = 0$. On a aussi $N_w = 3$ et $\tilde{N}_w = 2$.)

Pour un mot w, on note $\mathcal{C}_k$ l'ensemble des mots correspondant à un train comportant exactement k occurrences du mot w. Plus précisément, le mot $v = v_1 \cdots v_n$ est un élément de $\mathcal{C}_k$ si et seulement si

- le mot v commence et se termine par le mot w : $v_1 \cdots v_h = w$ et $v_{n-h+1} \cdots v_n = w$,
- pour la suite $(v_1, \dots, v_n)$, correspondant au mot v, on a $N_w = k$ et $\tilde{N}_w = 1$ (un seul train, mais k occurrences du mot w).

(Par exemple si $w = aba$, alors $\mathcal{C}_1 = \{aba\}$, $\mathcal{C}_2 = \{ababa\}$, $\mathcal{C}_3 = \{abababa\}$, ou encore si $w = abaaba$, alors $\mathcal{C}_1 = \{abaaba\}$, $\mathcal{C}_2 = \{abaabaaba, abaababaaba\}$.)

Notons que tout mot de $\mathcal{C}_k$ possède une longueur n comprise entre $h+k-1$ et $k(h-1)+1$. Enfin, on a $\mathcal{C}_1 = \{w\}$, et si le mot w ne peut pas se chevaucher lui-même, alors on a pour $k \geq 2$, $\mathcal{C}_k = \emptyset$.

Pour $k \geq 1$, on pose $\pi(\mathcal{C}_k) = \sum_{v \in \mathcal{C}_k} \pi(v)$, où la probabilité $\pi(v)$ est déterminée par (6.1), et on définit

$$\theta^{(k)}(w) = \pi(\mathcal{C}_k) - 2\pi(\mathcal{C}_{k+1}) + \pi(\mathcal{C}_{k+2}).$$

La quantité $\theta^{(k)}(w)$ s'interprète comme la probabilité, en régime stationnaire, d'observer un train de k occurrences exactement du mot w, débutant en position i donnée (avec i grand). En effet, on comprend intuitivement que si on observe une séquence v, qui est un train de k occurrences exactement du mot w, débutant en position i, cela signifie

- que la séquence v constitue un train comportant k occurrences du mot w : elle commence donc par un élément de $\mathcal{C}_k$ (ce qui arrive avec probabilité $\pi(\mathcal{C}_k)$),
- que la séquence v ne fait pas partie d'un train débutant en i et comportant au moins $k+1$ occurrences du mot w : la séquence débutant en i ne commence donc pas par un élément de $\mathcal{C}_{k+1}$ (ce qui arrive avec probabilité $\pi(\mathcal{C}_{k+1})$),
- que la séquence v ne fait pas partie d'un train qui débute avant la position i : elle ne représente donc pas la fin d'un train comportant au moins $k+1$ occurrences (ce qui arrive avec probabilité $\pi(\mathcal{C}_{k+1})$),
- enfin, dans les deux derniers événements, on a compté deux fois les trains comportant au moins $k+2$ occurrences qui débutent avant la position i et qui se terminent après la séquence v (ce qui arrive avec probabilité $\pi(\mathcal{C}_{k+2})$).

La probabilité, en régime stationnaire, d'observer un train débutant en position i donnée (avec i grand) est intuitivement donnée par

$$\theta(w) = \sum_{k\geq 1} \theta^{(k)}(w). \tag{6.8}$$

Remarquons que $\theta(w)$ est la probabilité d'observer un mot w débutant en position i, $\pi(w)$, moins la probabilité que ce mot appartienne à un train ayant commencé avant la position i. Ces trains comportent au moins deux occurrences, leur probabilité est donc $\pi(\mathcal{C}_2)$. On vérifie formellement que $\theta(w) = \pi(w) - \pi(\mathcal{C}_2)$:

$$\begin{aligned}\theta(w) &= \sum_{k\geq 1}[\pi(\mathcal{C}_k) - 2\pi(\mathcal{C}_{k+1}) + \pi(\mathcal{C}_{k+2})] \\ &= \sum_{k\geq 1}\pi(\mathcal{C}_k) - 2\sum_{k\geq 2}\pi(\mathcal{C}_k) + \sum_{k\geq 3}\pi(\mathcal{C}_k) = \pi(w) - \pi(\mathcal{C}_2).\end{aligned}$$

La probabilité, en régime stationnaire, d'observer un mot w débutant en position i, c'est-à-dire $\pi(w)$, peut se décomposer suivant le nombre d'occurrences du mot w dans le train auquel le mot w débutant en position i appartient et suivant sa position dans le train (k possibilités pour un train de k occurrences). On vérifie formellement que $\pi(w) = \sum_{k\geq 1} k\theta^{(k)}(w)$:

$$\begin{aligned}\sum_{k\geq 1} k\theta^{(k)}(w) &= \sum_{k\geq 1} k\pi(\mathcal{C}_k) - 2(k-1)\sum_{k\geq 2}\pi(\mathcal{C}_k) + (k-2)\sum_{k\geq 3}\pi(\mathcal{C}_k) \\ &= \pi(\mathcal{C}_1) = \pi(w).\end{aligned}$$

Le lemme suivant, dont nous aurons besoin par la suite, permet de justifier les calculs formels précédents. On pose

$$\alpha_w = \frac{\pi(\mathcal{C}_2)}{\pi(w)}. \tag{6.9}$$

Lemme 6.3.4. *On a $\alpha_w < 1$. Et pour $k \geq 1$, on a $\pi(\mathcal{C}_{k+1}) = \alpha_w \pi(\mathcal{C}_k)$, en particulier $\pi(\mathcal{C}_k) = \alpha_w^{k-1}\pi(w)$.*

On en déduit que $\theta^{(k)}(w) = (1-\alpha_w)^2\alpha_w^{k-1}\pi(w)$, puis les deux égalités suggérées ci-dessus :

$$\theta(w) = \sum_{k\geq 1} \theta^{(k)}(w) = (1-\alpha_w)\pi(w) = \pi(w) - \pi(\mathcal{C}_2),$$
$$\sum_{k\geq 1} k\theta^{(k)}(w) = \pi(w)(1-\alpha_w)^2 \sum_{k\geq 1} k\alpha_w^{k-1} = \pi(w).$$

Démonstration du lemme 6.3.4. Comme toutes les séquences de $\mathcal{C}_2$ commencent aussi par le mot w, on a $\pi(\mathcal{C}_2) \leq \pi(w)$. Nous démontrons par l'absurde que $\pi(\mathcal{C}_2) < \pi(w)$. On commence par démontrer que tous les trains de mots w sont p.s. finis. Soit w' un mot composé de h occurrences de la même lettre et distinct de w. En particulier, un train de mots w ne peut contenir le mot w'. Comme $\pi(w') = \pi(w'_1)P(w'_1, w'_1)^{h-1} > 0$, on déduit du théorème ergodique que p.s. $\lim_{N\to\infty} N_{w'}/N = \pi(w') > 0$. Ainsi, p.s. il existe $n \geq 1$, tel que $Y_n^{n+h} = w'$. Donc, tous les trains de w sont p.s. finis.

Si $\pi(\mathcal{C}_2) = \pi(w)$, cela signifie que p.s. toute séquence commençant par le mot w commence par un train comportant au moins deux occurrences. Mais la deuxième occurrence est alors p.s. aussi le début d'un train comportant au moins deux occurrences. En itérant ce raisonnement, on obtient qu'une séquence commençant par le mot w est p.s. une succession de mots w se chevauchant. Et donc les trains de w sont p.s. infinis, ce qui contredit le raisonnement précédent. Donc, on a $\pi(\mathcal{C}_2) < \pi(w)$.

On suppose $k \geq 2$. Nous vérifions maintenant qu'il existe une bijection, φ, entre $\mathcal{C}_{k+1}$ et $\mathcal{C}_2 \times \mathcal{C}_k$. Si $v \in \mathcal{C}_{k+1}$, alors le mot v commence par un mot $u \in \mathcal{C}_2$. On définit le mot z, tel que v soit la concaténation de u et z : $v = uz$. Comme $v \in \mathcal{C}_{k+1}$ et que u est un train comportant seulement deux occurrences du mot w, on en déduit que le mot wz est un train comportant exactement k occurrences du mot w. On définit $\varphi(v) = (u, wz)$. Par construction φ est une injection de $\mathcal{C}_{k+1}$ dans $\mathcal{C}_2 \times \mathcal{C}_k$. Pour tout couple $(u, y) \in \mathcal{C}_2 \times \mathcal{C}_k$, comme le mot y commence par le mot w, on peut définir z, tel que $y = wz$, et considérer le mot $v = uz$. Par construction v est un train comportant $k+1$ occurrences du mot w. Donc on a $v \in \mathcal{C}_{k+1}$ ainsi que $\varphi(v) = (u, wz) = (u, y)$. La fonction φ est une surjection, donc une bijection. Rappelons que si $u \in \mathcal{C}_2$ alors la dernière lettre de u est w_h et si $y \in \mathcal{C}_k$, alors y est de la forme wz et on a $\pi(y) = \pi(w)P(w_h, z_1)\pi(z)/\pi(z_1)$. Remarquons ensuite que pour $u \in \mathcal{C}_2$, $y = wz \in \mathcal{C}_k$, on a $\varphi^{-1}(u, y) = uz$ et

$$\pi(\varphi^{-1}(u, y)) = \pi(uz) = \frac{\pi(u)P(w_h, z_1)\pi(z)}{\pi(z_1)} = \frac{\pi(u)\pi(y)}{\pi(w)}.$$

On en déduit donc que pour $k \geq 2$,

$$\pi(\mathcal{C}_{k+1}) = \sum_{v \in \mathcal{C}_{k+1}} \pi(v) = \sum_{u \in \mathcal{C}_2, y \in \mathcal{C}_k} \pi(\varphi^{-1}(u,y))$$
$$= \sum_{u \in \mathcal{C}_2, y \in \mathcal{C}_k} \frac{\pi(u)\pi(y)}{\pi(w)} = \frac{\pi(\mathcal{C}_2)\pi(\mathcal{C}_k)}{\pi(w)} = \alpha_w \pi(\mathcal{C}_k).$$

□

Dans ce qui suit, nous présentons sans démonstration les théorèmes asymptotiques, et nous renvoyons à [11] pour un exposé complet de ces résultats.

On considère une suite de mots $(w_N, N \geq 3)$ de longueur h_N telle que $N\pi(w_N) = O(1)$. Ceci est automatiquement réalisé, si $h_N \geq 1 + c \log N$, pour une constante c assez grande. En effet, comme $m = \max\{P(y,y'), y, y' \in E\} \in]0,1[$, on a alors $\pi(w_N) \leq m^{h_N - 1}$ et $N\pi(w_N) \leq \exp[(h_N - 1)\log(m) + \log(N)]$. En particulier on a $N\pi(w_N) \leq 1$ dès que $c \geq 1/\log(1/m)$.

On note $\mu_{\tilde{N}_w} = (\mu_{\tilde{N}_w}(k) = \mathbb{P}(\tilde{N}_w = k), k \in \mathbb{N})$ la loi de $\tilde{N}_w$, le nombre de trains de mots w dans une séquence de longueur N. On note également $\rho_\theta = (\rho_\theta(k) = \mathrm{e}^{-\theta}\,\theta^k/k!, k \in \mathbb{N})$ la loi de Poisson de paramètre θ. On a le résultat suivant sur la convergence au sens de la norme en variation de la loi de $\tilde{N}_{w_N}$ vers une loi de Poisson.

Proposition 6.3.5. *Soit $(w_N, N \geq 3)$ une suite de mots de longueur h_N telle que $N\pi(w_N) = O(1)$ et $h_N = o(N)$. Alors, on a*

$$\lim_{N \to \infty} \left\| \mu_{\tilde{N}_{w_N}} - \rho_{N\theta(w_N)} \right\| = 0.$$

Si le mot w ne peut pas se chevaucher lui-même, alors $\tilde{N}_w = N_w$, et la loi de N_w est donc proche (au sens de la norme en variation) asymptotiquement de la loi de Poisson de paramètre $N\theta(w) = N(\pi(w) - \pi(\mathcal{C}_2)) = N\pi(w)$.

Enfin, si les mots se chevauchent, la description de la loi asymptotique de N_w est plus complexe. Pour cela, on considère $\tilde{N}_w^{(k)}$ le nombre d'occurrences de trains comportant exactement k mots w se chevauchant. En particulier, on a $\tilde{N}_w = \sum_{k \geq 1} \tilde{N}_w^{(k)}$ et $N_w = \sum_{k \geq 1} k\tilde{N}_w^{(k)}$. Sous les hypothèses du théorème précédent, on peut vérifier que $(\tilde{N}_{w_N}^{(k)}, k \geq 1)$ se comporte asymptotiquement comme $(V_k, k \geq 1)$, où les variables aléatoires $(V_k, k \geq 1)$ sont indépendantes et la loi de V_k est la loi de Poisson de paramètres $N\theta^{(k)}(w_N)$. Ainsi les trains comportant exactement k occurrences du mot w suivent asymptotiquement une « loi des petits nombres ». De plus, les occurrences, correspondant à des trains ne comportant pas le même nombre de fois le mot w, sont asymptotiquement indépendantes. Cela permet alors de démontrer le résultat suivant.

Proposition 6.3.6. *Soit $(w_N, N \geq 3)$ une suite de mots de longueur h_N telle que $N\pi(w_N) = O(1)$ et $h_N = o(N)$. Alors, on a*

$$\lim_{N \to \infty} \left\| \mu_{N_{w_N}} - \nu_N \right\| = 0,$$

où ν_N est la loi de $\sum_{k\geq 1} kV_k^N$, les variables $(V_k^N, k \geq 1)$ étant indépendantes et distribuées suivant les lois de Poisson de paramètres respectifs $N\theta^{(k)}(w_N)$.

Bien sûr les paramètres $\theta^{(k)}(w_N)$ sont inconnus. Mais ils peuvent être estimés par les estimateurs convergents suivants :

$$\hat{\theta}_N^{(k)}(w_N) = (1-\hat{\alpha}_{w_N})^2 \hat{\alpha}_{w_N}^{k-1} \hat{\pi}_N(w_N), \quad \text{avec} \quad \hat{\alpha}_{w_N} = \frac{\sum_{v\in\mathcal{C}_2(w_N)} \hat{\pi}_N(v)}{\hat{\pi}_N(w_N)},$$

où $\hat{\pi}_N(v)$ est l'estimateur convergent de $\pi(v)$ donné par (6.5), et $\mathcal{C}_2(w_N)$ est l'ensemble $\mathcal{C}_2$ défini pour le mot w_N : c'est l'ensemble des trains comportant exactement 2 occurrences du mot w_N. On peut alors montrer (voir [11]) que la convergence établie dans la proposition 6.3.6 reste valide si l'on remplace les paramètres par leur estimation. Plus précisément, on a le résultat suivant.

Théorème 6.3.7. *Soit $(w_N, N \geq 3)$ une suite de mots de longueur h_N telle que $N\pi(w_N) = O(1)$ et $h_N = o(N)$. Alors, on a*

$$\lim_{N\to\infty} \left\| \mu_{N_{w_N}} - \tilde{\nu}_N \right\| = 0,$$

où $\tilde{\nu}_N$ est la loi de $\sum_{k\geq 1} k\tilde{V}_k^N$, les variables $(\tilde{V}_k^N, k \geq 1)$ étant indépendantes et distribuées suivant les lois de Poisson de paramètres respectifs $N\hat{\theta}_N^{(k)}(w_N)$.

On peut alors reproduire un raisonnement similaire à celui développé après le corollaire 6.1.2 pour détecter les mots exceptionnels.

6.4 Un autre modèle pour la séquence d'ADN

À la fin du paragraphe 6.2, nous avons souligné le fait que dans le modèle de chaîne de Markov considéré, il n'est pas cohérent d'utiliser le nombre d'occurrences de $w-$ quand on cherche à détecter les mots exceptionnels (voir aussi le troisième point de la remarque 6.1.3) en dehors du cas où w est un mot de longueur 3. Ceci remet en cause l'approche élémentaire du paragraphe 6.1. En fait, si l'on considère des modèles de chaînes de Markov d'ordre $h-2$, où h est la longueur du mot w, voir la définition ci-dessous, alors le nombre d'occurrences de $w-$ est dans la statistique exhaustive et apparaît naturellement dans la construction d'estimateurs du maximum de vraisemblance. Ceci suggère que l'approche du paragraphe 6.1 s'inscrit plutôt dans le cadre d'un modèle de chaîne de Markov d'ordre $h-2$.

L'objectif de ce paragraphe est de généraliser le théorème 6.1.1 et le corollaire 6.1.2 au modèle de chaîne de Markov d'ordre supérieur, en utilisant des techniques similaires. De ce fait, les résultats seront suggérés au travers d'un problème.

Définition 6.4.1. *On dit que la suite $Y = (Y_n, n \geq 1)$ est une chaîne de Markov d'ordre m, si la suite $(Y_{n-m+1}^n, n \geq m)$ est une chaîne de Markov.*

Remarquons qu'une chaîne de Markov est une chaîne de Markov d'ordre 1, et que si Y est une chaîne de Markov d'ordre m, alors elle est également d'ordre $k \geq m$.

Problème 6.4.2. On considère $Y = (Y_n, n \geq 1)$ une chaîne de Markov d'ordre $m \geq 1$ sur E, fini non réduit à un singleton. Pour tous $y \in E^m, z \in E$, on pose

$$Q(y\,;z) = \mathbb{P}(Y_{m+1} = z | Y_1^m = y),$$

et on suppose que $Q(y\,;z) \in]0,1[$.

1. Vérifier que la matrice de transition de la chaîne $\tilde{Y} = (Y_{n-m+1}^n, n \geq m)$ est définie de la manière suivante : pour $y = y_1^m$, $y' = {y'}_1^m \in E^m$,

$$P(y,y') = \mathbf{1}_{\{{y'}_1^{m-1} = y_2^m\}} Q(y\,;y'_m).$$

En déduire que la chaîne de Markov $\tilde{Y}$ est irréductible.

On note $\pi = (\pi(y), y \in E^m)$ la probabilité invariante de $\tilde{Y}$.

Pour un mot $v = v_1 \cdots v_l$ de longueur l, le nombre d'occurrences du mot v dans une séquence de longueur N est défini par

$$N_v = \sum_{k=l}^{N} \mathbf{1}_{\{Y_{k-l+1}^k = v\}},$$

et si $l \geq m+1$, on pose

$$\pi(v) = \pi(v_1^m) \prod_{i=1}^{l-m} Q(v_i^{i+m-1}\,; v_{i+m}).$$

Soit $w = w_1 \cdots w_h$ un mot de longueur $h = m+2$. En particulier, on a $\pi(w) = \pi(w_1^{h-2}) Q(w_1^{h-2}\,; w_{h-1}) Q(w_2^{h-1}\,; w_h)$.

2. Pour $y \in E^m$ et $z \in E$, on pose $\hat{Q}_N(y\,;z) = N_{yz}/N_y$. Montrer que $\hat{Q}_N(y\,;z)$ est un estimateur convergent de $Q(y\,;z)$. On pourra également vérifier, en s'inspirant de la démonstration du lemme 6.2.1, que $\hat{Q}_N(y\,;z)$ est un estimateur asymptotiquement normal de même variance asymptotique que l'estimateur du maximum de vraisemblance de $Q(y\,;z)$.
3. Vérifier que, pour tout $y \in E^m$, $\hat{\pi}_N(y) = N_y/N$ est un estimateur convergent de $\pi(y)$.
4. Déduire des questions précédentes un estimateur de $\pi(w)$ construit à partir des estimateurs $(\hat{Q}_N(y\,;z), y \in E^m, z \in E)$ et $\hat{\pi}_N = (\hat{\pi}_N(y), y \in E^m)$.
5. Montrer que p.s. $\lim_{N\to\infty} N_w/N = \pi(w)$.
6. On rappelle que $m = h-2$. Vérifier que $(X_n = Y_{n-h+2}^n, n \geq h-1)$ est une chaîne de Markov irréductible sur E^{h-1}, de matrice de transition définie pour $x = x_1^{h-1}$, $x' = {x'}_1^{h-1} \in E^{h-1}$ par

$$P^X(x,x') = \mathbf{1}_{\{{x'}_1^{h-2} = x_2^{h-1}\}} Q({x'}_1^{h-2}\,; x'_{h-1}),$$

et de probabilité invariante

$$\pi^X(x_1^{h-1}) = \pi(x_1^{h-2})Q(x_1^{h-2}\,;x_{h-1}).$$

On note $w- = w_1\cdots w_{h-1}$ le mot w tronqué de sa dernière lettre, $-w = w_2\cdots w_h$, le mot w tronqué de sa première lettre et $-w- = w_2\cdots w_{h-1}$, le mot w tronqué de ses première et dernière lettre. On a

$$\pi(w-) = \pi(w_1^{h-2})Q(w_1^{h-2}\,;w_{h-1}) \quad \text{et} \quad \pi(-w) = \pi(-w-)Q(-w-\,;w_{h-1}).$$

On considère les fonctions définies pour $x = x_1^{h-1}, x' = {x'}_1^{h-1} \in E^{h-1}$:

$$g_1(x,x') = \mathbf{1}_{\{x_1^{h-1}=w-,x'_{h-1}=w_h\}} \quad \text{et} \quad g_2(x,x') = \mathbf{1}_{\{x_2^{h-1}=-w-,x'_{h-1}=w_h\}}.$$

7. Vérifier que

$$P^X g_1(x_1^{h-1}) = \mathbf{1}_{\{x_1^{h-1}=w-\}}Q(-w-\,;w_h),$$
$$P^X g_2(x_1^{h-1}) = \mathbf{1}_{\{x_2^{h-1}=-w-\}}Q(-w-\,;w_h)$$

ainsi que $(\pi^X, P^X g_1) = \pi(w)$, $(\pi^X, (P^X g_1)^2) = \pi(w)Q(-w-\,;w_h)$, et, en utilisant le fait que π est la probabilité invariante associée à P, que $(\pi^X, P^X g_2) = \pi(-w)$ et $(\pi^X, (P^X g_2)^2) = \pi(-w)Q(-w-\,;w_h)$.

8. Calculer $\sum_{k=h}^N g_1(X_{k-1},X_k)$, $\sum_{k=h}^N P^X g_1(X_{k-1})$, $\sum_{k=h}^N g_2(X_{k-1},X_k)$, $\sum_{k=h}^N P^X g_2(X_{k-1})$.

On pose

$$\zeta''_N = \frac{1}{\sqrt{N}}\left(N_w - \frac{N_{w-}N_{-w}}{N_{-w-}}\right).$$

9. Montrer en s'inspirant de la démonstration du théorème 6.1.1 que la suite $(\zeta''_N, N \geq h)$ converge en loi vers une variable gaussienne centrée de variance

$$\sigma''^2 = \pi(w)\Big[1 - \frac{\pi(w-)}{\pi(-w-)}\Big][1 - Q(-w-\,;w_h)].$$

On pose

$$Z''_N = \zeta''_N/\sigma''_N, \tag{6.10}$$

où $\sigma''_N = \sqrt{{\sigma''_N}^2}$ et ${\sigma''_N}^2 = \dfrac{N_{w-}N_{-w}}{NN_{-w-}}\Big[1 - \dfrac{N_{w-}}{N_{-w-}}\Big]\Big[1 - \dfrac{N_{-w}}{N_{-w-}}\Big]$.

10. Donner l'analogue du corollaire 6.1.2 pour un modèle de chaîne de Markov d'ordre $m = h-2$. En déduire un test pour identifier les mots exceptionnels construits à partir de la statistique Z''_N.

Les quantités N_{w-}, N_{-w} apparaissent naturellement dans l'estimation de $\pi(w)$ à l'aide de l'estimateur du maximum de vraisemblance dans un modèle de chaîne de Markov d'ordre $h-2$. ⧫

6.5 Conclusion

D'après les commentaires qui suivent le lemme 6.2.1, on désire connaître la loi du nombre exact d'occurrences d'un mot w connaissant le nombre d'occurrences des mots de longueur deux, ainsi que la valeur de la première lettre (i.e. y_1). Elle est en fait connue explicitement mais difficile à calculer numériquement. Elle permet toutefois de vérifier la validité des théorèmes limites présentés dans ce chapitre (voir [10]). L'approximation par la « loi des petits nombres » semble mieux se comporter pour les mots ayant un faible nombre d'occurrences (mots rares ou mots longs). Il faut également citer l'existence de méthodes de type grandes déviations (voir [7]) pour tester si des groupes de mots sont exceptionnellement rares ou fréquents.

Dans l'exemple 6.2.5 nous avons observé sur une simulation le comportement des variables Z_N, définies au paragraphe 6.1, et Z'_N, définies au paragraphe 6.2, pour une matrice de transition proche de celle estimée pour E. coli (voir ci-dessous).

Les résultats numériques qui suivent concernent la séquence circulaire de l'ADN de E. coli, extraite de [4]. La matrice P estimée par la formule du lemme 6.2.1 est

$$\hat{P}_N \simeq \begin{pmatrix} 0.29579 & 0.22472 & 0.20825 & 0.27124 \\ 0.27566 & 0.23032 & 0.29390 & 0.20012 \\ 0.22709 & 0.32620 & 0.22951 & 0.21720 \\ 0.18578 & 0.23426 & 0.28242 & 0.29755 \end{pmatrix}. \tag{6.11}$$

Les figures présentent toutes les données, sauf certaines ayant des valeurs extrêmes (voir le Tableau 6.1 pour les valeurs maximales et minimales des statistiques). On donne dans la Fig. 6.5 (resp. 6.8) les histogrammes des variables Z_N et Z'_N calculées pour tous les mots de longueur 6 (resp. 8). On remarquera que ni les grandeurs typiques (entre -50 et 50 pour les mots de longueur 6 et entre -15 et 15 pour les mots de longueur 8) ni l'allure des histogrammes ne correspondent à des réalisations de variables aléatoires gaussiennes centrées réduites. Les figures sont très différentes de celles obtenues dans l'exemple 6.2.5 sur des simulations. Ainsi, le modèle de chaîne de Markov vu au paragraphe 6.1 et 6.2 semble inadapté. En particulier, les données de l'ADN d'E. coli, ne peuvent raisonnablement pas être modélisées par une chaîne de Markov. En fait des modélisations plus complexes de l'ADN, prenant en compte plusieurs phénomènes biologiques, sont actuellement utilisées pour l'analyse statistique de l'ADN.

Pour les mots de longueur 6 (resp. 8), la Fig. 6.6 (resp. 6.9) présente les points de coordonnées (Z_N, Z'_N) ; la Fig. 6.7 (resp. 6.10) présente les points de coordonnées (Z_N, Z''_N) et (Z'_N, Z''_N), où les variables Z''_N sont définies au paragraphe 6.4 par (6.10) (il s'agit d'un modèle de chaîne de Markov d'ordre 4 pour les mots de longueur 6, et d'ordre 6 pour les mots de longueur 8). On remarque également que le nuage de points (Z_N, Z'_N) est globalement situé autour de la diagonale, alors que l'on observe un comportement différent pour

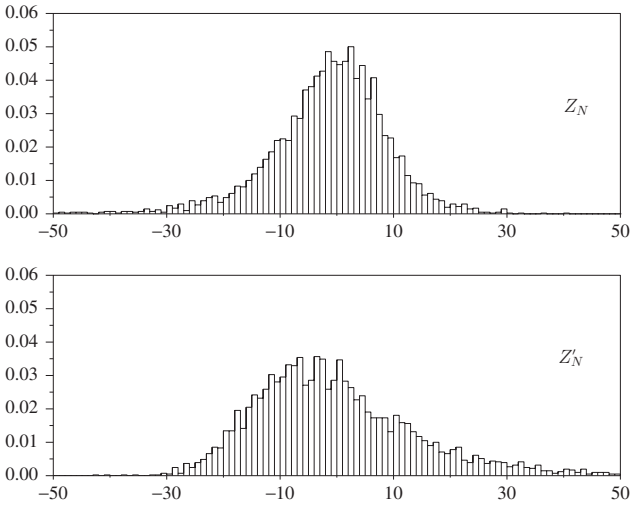

Fig. 6.5. Histogrammes de Z_N et Z'_N pour les mots de longueur 6 observés sur l'ADN de E. coli

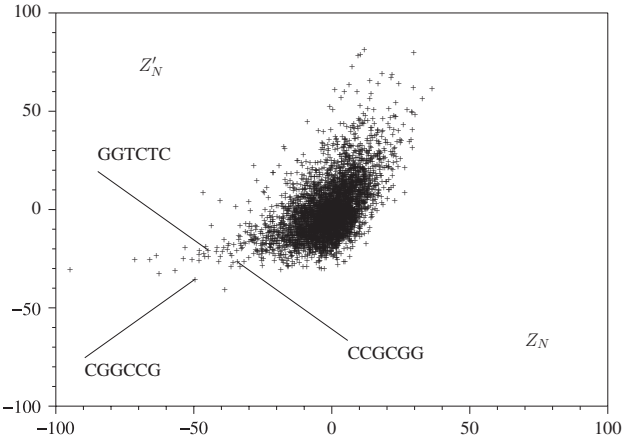

Fig. 6.6. (Z_N, Z'_N) pour les mots de longueur 6 observés sur l'ADN de E. coli

les couples (Z_N, Z''_N) et (Z'_N, Z''_N). Les résultats sont donc sensibles à l'ordre de la chaîne de Markov, ce qui souligne encore une fois que le modèle de chaîne de Markov (d'ordre 1) n'est pas adapté.

Enfin, nous présentons, dans le Tableau 6.2, les résultats numériques obtenus pour les trois sites de restriction et la séquence Chi présentés en introduction de ce chapitre :

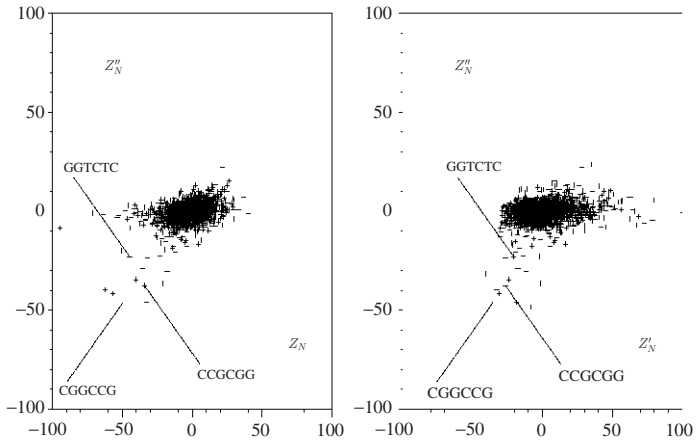

Fig. 6.7. (Z_N, Z''_N) (figure de gauche) et (Z'_N, Z''_N) (figure de droite) pour les mots de longueur 6 observés sur l'ADN de E. coli

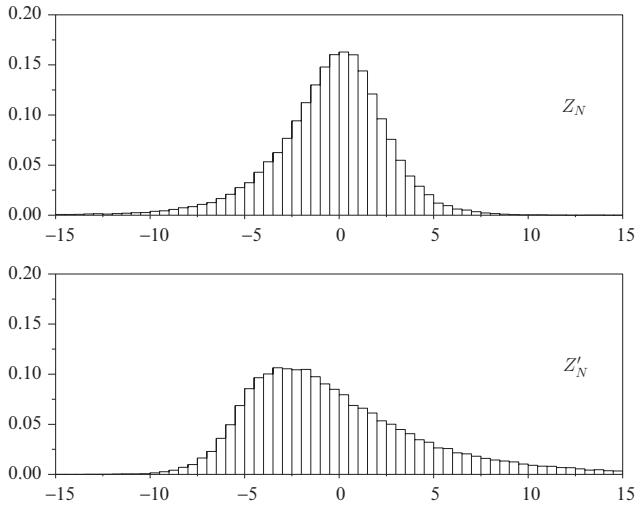

Fig. 6.8. Histogrammes de Z_N et Z'_N pour les mots de longueur 8 observés sur l'ADN de E. coli

- La première ligne indique le nombre d'occurrences des séquences concernées.
- Pour le modèle développé au paragraphe 6.1, on donne le nombre d'occurrences attendu estimé, $N_{w-}N_{w_{h-1}w_h}/N_{w_{h-1}}$, la statistique de test Z_N, ainsi que son rang parmi tous les Z_N des mots de même longueur ($4^6 = 4\,096$ mots distincts pour les sites de restriction considérés qui

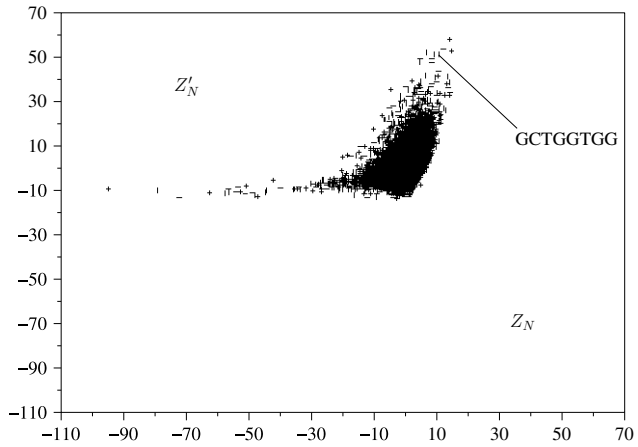

Fig. 6.9. (Z_N, Z'_N) pour tous les mots de longueur 8 observés sur l'ADN de E. coli

Tableau 6.1. Valeurs extrêmales des statistiques pour les mots de longueur 6 et 8

Pour une simulation, voir l'exemple 6.2.5.

Longueur des mots	$\min(Z_N)$	$\max(Z_N)$	$\min(Z'_N)$	$\max(Z'_N)$	$\min(Z''_N)$	$\max(Z''_N)$
6	- 3.8	3.6	- 3.5	3.9	- 3.8	3.7
8	- 5.3	3.7	- 4.1	4.6	- 5.8	3.5

Pour la séquence de E. coli.

Longueur des mots	$\min(Z_N)$	$\max(Z_N)$	$\min(Z'_N)$	$\max(Z'_N)$	$\min(Z''_N)$	$\max(Z''_N)$
6	-218.4	40.2	-42.6	101.7	-267.9	24.0
8	- 94.7	14.4	- 13.5	58.0	- 21.0	8.2

comportent 6 lettres, et $4^8 = 65\,536$ mots distincts pour le motif Chi de 8 lettres).

– Pour le modèle développé au paragraphe 6.2, on donne le nombre d'occurrences attendu estimé, $\hat{\pi}_N(w)$, la statistique de test Z'_N, ainsi que son rang parmi tous les Z'_N des mots de même longueur.
– Pour le modèle développé au paragraphe 6.4 (variante du premier modèle pour des chaînes de Markov d'ordre supérieur), on donne le nombre d'occurrences attendu estimé, $N_{w-}N_{-w}/N_{-w-}$, la statistique de test Z''_N, ainsi que son rang parmi tous les Z''_N des mots de même longueur.

Enfin, les résultats numériques correspondant à l'approximation par la « loi des petits nombres », présentée au paragraphe 6.3, nécessitent des calculs plus délicats pour les mots pouvant se chevaucher. Ils ne sont pas présentés ici, mais ils sont disponibles grâce aux logiciels de l'INRA (voir le site `http://www-mig.jouy.inra.fr/ssb/rmes/`).

Pour les modèles ci-dessus, presque tous les mots ont des p-valeurs extrêmement faibles ou extrêmement élevées (on a déjà remarqué que le modèle

Tableau 6.2. Valeurs calculées pour trois sites de restriction de longueur 6 (4 096 mots de longueur 6), et pour le motif Chi de longueur 8 (65 536 mots de longueur 8) de E. coli

Séquence	GGTCTC	CGGCCG	CCGCGG	GCTGGTGG
Nombre d'occurrences (N_w)	124	284	657	499
Modèle du paragraphe 6.1				
$N_{w-}N_{w_{h-1}w_h}/N_{w_{h-1}}$	559.9	984.0	1425.3	291.9
Z_N	-44.8	-49.5	-34.2	10.6
rang	4077	4085	4055	31
rang (%)	99.5 %	99.7 %	99.0 %	0.05 %
Modèle du paragraphe 6.2				
$\hat{\pi}_N(w)$	644.2	1756.7	1756.7	70.1
Z'_N	-20.6	-35.4	-26.5	51.3
rang	3936	4093	4069	6
rang (%)	96.1 %	99.9 %	99.3 %	0.01 %
Modèle du paragraphe 6.4				
$N_{w-}N_{-w}/N_{-w-}$	332.0	859.2	1404.7	420.2
Z''_N	-23.0	-46.2	-37.6	5.7
rang	4079	4093	4089	27
rang (%)	99.6 %	99.9 %	99.8 %	0.04 %

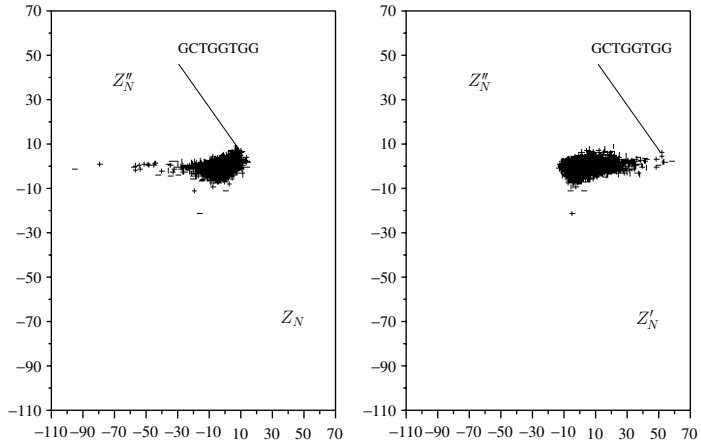

Fig. 6.10. (Z_N, Z''_N) (figure de gauche) et (Z'_N, Z''_N) (figure de droite) pour tous les mots de longueur 8 observés sur l'ADN de E. coli

n'était pas vraiment adapté aux observations). En revanche, si l'on ne peut pas appliquer les procédures de tests décrites dans les paragraphes précédents, il est intéressant de regarder les rangs des statistiques. En particulier, on constate dans le Tableau 6.2 que les trois sites de restriction ont des rangs très

élevés dans les trois modèles (i.e. des valeurs très négatives des statistiques, parmi les cinquante dernières sur $4^6 = 4\,096$, sauf pour une valeur), et le motif Chi a un rang très faible (i.e. des valeurs très positives des statistiques, parmi les quarante premiers mots sur $4^8 = 65\,536$). On pourra comparer les positions des mots d'intérêt, sites de restriction et motif Chi, dans les nuages de points correspondant à une simulation (Figs. 6.3 et 6.4) et les nuages correspondant à l'ADN d'E. coli (Figs. 6.5 et 6.8). En ce sens le modèle détecte bien ces mots qui possèdent un rôle biologique.

Références

1. D. Aldous. *Probability approximations via the Poisson clumping heuristic*, volume 77 de *Applied Mathematical Sciences*. Springer-Verlag, 1989.
2. A. Barbour, L. Holst et S. Janson. *Poisson approximation*. Oxford Studies in Probability. Clarendon Press, 1992.
3. P. Bickel et K. Doksum. *Mathematical statistics. Basic ideas and selected topics*. Holden-Day Series in Probability and Statistics. Holden-Day, San Francisco, 1977.
4. F. Blattner, G. Plunkett, C. Bloch, N. Perna, V. Burland, M. Riley, J. Collado-Vides, J. Glasner, C. Rode, G. Mayhew, J. Gregor, N. Davis, H. Kirkpatrick, M. Goeden, D. Rose, B. Mau et Y. Shao. The complete genome sequence of Escherichia coli K-12. *Science*, 277 : 1453–1474, 1997.
5. M. El Karoui, M. Schaeffer, V. Biaudet, A. Bolotin, A. Sorokin et A. Gruss. Orientation specificity of the Lactococcus lactis Chi site. *Genes to Cells*, 5 : 453–461, 2000.
6. S. Kowalczykowski, D. Dixon, A. Eggleston, S. Lauder et W. Rehauer. Biochemistry of homologous recombination in Escherichia coli. *Microbiol. Rev.*, 58 : 401–465, 1994.
7. G. Nuel. *Grandes déviations et chaînes de Markov pour l'étude des occurrences de mots dans les séquences biologiques*. Thèse, Université d'Évry Val d'Essonne, 2001.
8. B. Prum, F. Rodolphe et E. de Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J.R. Statist. Soc. B*, 57(1) : 205–220, 1995.
9. S. Robin, F. Rodolphe et S. Schbath. *ADN, mots et modèles*. Belin, 2003.
10. S. Robin et S. Schbath. Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.*, 2001.
11. S. Schbath. *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquences exceptionnelles dans les séquences d'ADN*. Thèse, Université René Descartes (Paris V), 1995.
12. E. Seneta. *Nonnegative matrices and Markov chains*. Springer, New-York, seconde édition, 1981.