Alignement de séquences : modèles stochastiques

Aimé Lachal

Mai 2001

Le problème de l'alignement de séquences d'ADN en génomique consiste à comparer une séquence d'ADN prélevée sur un individu à une séquence stockée dans une base de données dont les propriétés sont déjà connues. La modélisation aléatoire vise à représenter une séquence biologique par une suite de v.a. $(X_n)_{n\geqslant 1}$. De nombreux modèles ont déjà été introduits : modèle de Bernoulli (les X_n sont indépendantes de même loi de Bernoulli) et modèles markoviens (dépendance faible : la base X_n ne dépend que de la précédente, X_{n-1}, dans la séquence) pour les ADN et ARN ou modèles markoviens d'ordre p (dépendance plus forte : la base X_n dépend de $X_{n-1}, X_{n-2}, \ldots, X_{n-p}$) pour les protéines. Les modèles markoviens cachés prennent en considération des états cachés correspondant à des parties codantes (exons) ou non (introns) lors de la traduction des ARN en protéines.

1 Critères de similarité

Disposant de deux séquences, la séquence observée (séquence requête) $\boldsymbol{A} = a_1 \ldots a_m$ et une séquence $\boldsymbol{B} = b_1 \ldots b_n$ de la banque de données, les a_i et b_j appartenant à un alphabet $\mathcal{A}$, la similarité de ces deux séquences s'étudie selon la comparaison entre chaque a_i et b_i. Un alignement des séquences $\boldsymbol{A}$ et $\boldsymbol{B}$ est une superposition de celles-ci obtenue en translatant éventuellement des lettres (on insère ainsi des trous appelés insertions-délétions ou indels notés $-$) de façon à obtenir un maximum d'identités entre les lettres superposées. Plus généralement, on envisage toute les suites que l'on peut obtenir à partir des suites $\boldsymbol{A}$ et $\boldsymbol{B}$ en insérant des trous correspondant aux indels de manière à ce qu'elles soient de même longueur l. Soit alors $\boldsymbol{A}^* = a_1^* \ldots a_l^*$ et $\boldsymbol{B}^* = b_1^* \ldots b_l^*$ les suites ainsi obtenues, $a_i^*, b_i^* \in \mathcal{A}^* = (\mathcal{A} \times \mathcal{A}) \cup \{(a,-),(-,b), a, b \in \mathcal{A}\}$. On interdit que deux trous se retrouvent au même indice dans les deux séquences $\boldsymbol{A}^*$ et $\boldsymbol{B}^*$. Alors $m \wedge n \leqslant l \leqslant m+n$. Chaque couple $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ est un alignement global possible de $(\boldsymbol{A}, \boldsymbol{B})$. On dira que la comparaison $a_i^* \longleftrightarrow b_i^*$ est une identité (resp. substitution, indel) si $a_i^* = b_i^*$ (resp. $a_i^* \neq b_i^*$, a_i^* ou $b_i^* = -$). La qualité d'un alignement $\binom{a_1^*}{b_1^*} \cdots \binom{a_l^*}{b_l^*}$ se teste selon différents critères, déterministes ou non.

1.1 Plus long segment commun

Le plus long segment commun est défini par

$$R(\boldsymbol{A}, \boldsymbol{B}) = \max\{k \in \{1, \ldots, m \wedge n\} : \exists (i,j), \forall l \in \{0, \ldots, k-1\}, a_{i+l} = b_{j+l}\}.$$

Par exemple, pour $\boldsymbol{A} = AAGTTC$ et $\boldsymbol{B} = AGCCC$, le plus long segment commun est AG et $R(\boldsymbol{A}, \boldsymbol{B}) = 2$. Pour être un peu moins restrictif, on introduit également le plus long segment

commun de qualité α :

$$R^{\alpha}(\boldsymbol{A},\boldsymbol{B}) = \max\{k \in \{1,\dots,m\wedge n\} : \exists (i,j), \sum_{l=0}^{k-1} \mathbb{1}_{\{a_{i+l}=b_{j+l}\}} \geqslant k\alpha\}.$$

Par exemple, pour $\boldsymbol{A} = AAGTTC$, $\boldsymbol{B} = AGCCC$ et $\alpha = \frac{1}{3}$, les plus longs segments communs de qualité α sont $AGTTC$ et $AGCCC$ et $R^{\alpha}(\boldsymbol{A},\boldsymbol{B}) = 5$.

1.2 Distance de Levensthein (ou distance d'édition)

La distance de Levensthein $D(\boldsymbol{A},\boldsymbol{B})$ entre les séquences $\boldsymbol{A}$ et $\boldsymbol{B}$ est le nombre minimal d'opérations élémentaires (substitutions, insertions, délétions) transformant $\boldsymbol{A}$ en $\boldsymbol{B}$. Par exemple, pour $\boldsymbol{A} = AAGTC$ et $\boldsymbol{B} = AGCC$, la succession suivante

$$\begin{array}{rcccccc} \boldsymbol{A} & = & A & A & G & T & C \\ & & - & A & G & T & C \\ & & - & A & G & C & C & = & \boldsymbol{B} \end{array}$$

donne $D(\boldsymbol{A},\boldsymbol{B}) = 2$.

1.3 Score d'alignement

L'utilisation d'une fonction de score $s : \mathcal{A}^* \longrightarrow \mathbb{R}$ permet de calculer un score relatif à la suite finie $(A_1^*, B_1^*), \dots, (A_l^*, B_l^*)$ que l'on cherchera à maximiser en vue de tester la pertinence de l'alignement. Pour cela on pose $\sigma(\boldsymbol{A}^*,\boldsymbol{B}^*) = \sum_{i=1}^{l} s(a_i^*, b_i^*)$ et l'on évalue

1. le score global : $S(\boldsymbol{A},\boldsymbol{B}) = \max\{\sigma(\boldsymbol{A}^*,\boldsymbol{B}^*), (\boldsymbol{A}^*,\boldsymbol{B}^*) \text{ alignement de } (\boldsymbol{A},\boldsymbol{B})\}$;
2. le score local : $H(\boldsymbol{A},\boldsymbol{B}) = \max\{S(I,J), I \subset \boldsymbol{A}, J \subset \boldsymbol{B}\}$.

Un exemple de fonction de score est donné par

$$s(a_i^*, b_i^*) = \begin{cases} 1 & \text{si } a_i^* \longleftrightarrow b_i^* \text{ est une identité,} \\ \alpha & \text{si } a_i^* \longleftrightarrow b_i^* \text{ est une substitution,} \\ \beta & \text{si } a_i^* \longleftrightarrow b_i^* \text{ est une insertion-délétion,} \end{cases}$$

les paramètres $\alpha < 0$ et $\beta < 0$ étant des pénalités de substitution et d'insertion-délétion. Le score d'alignement vaut ici $\sigma(\boldsymbol{A}^*,\boldsymbol{B}^*) = n_1 + \alpha n_2 + \beta n_3$ où n_1 (resp. n_2, n_3) est le nombre d'identités (resp. de substitutions, d'insertions-délétions).

Dans le cas d'alignement où les insertions-délétions ne sont pas admises, on compare deux séquences de même longueur. Les scores $S(\boldsymbol{A},\boldsymbol{B})$ et $H(\boldsymbol{A},\boldsymbol{B})$ se réduisent respectivement à $\sigma(\boldsymbol{A},\boldsymbol{B})$ et $\max\{\sigma(I,J), I \subset \boldsymbol{A}, J \subset \boldsymbol{B}, I, J \text{ de même longueur}\}$, ce qui revient à considérer les scores $S(X)$ et $H(X)$ pour des séquences définies sur l'alphabet $\mathcal{A} \times \mathcal{A}$. Ce cas de figure sera adopté dans toute la partie modélisation stochastique ci-dessous.

Signalons enfin qu'une fonction de score définie sur l'alphabet originel $\mathcal{A}$ peut être employée dans un contexte autre que celui de l'alignement, en vue de détecter par exemple une région présentant des propriétés particulières ou anormales : région codante (exon), non-codante (intron), amphotère, hydrophobe... On privilégiera cet aspect dans le paragraphe modélisation stochastique ci-dessous.

1.4 Algorithmes d'alignement

Dans cette partie, on ne présente que deux algorithmes d'alignement historiques : l'algorithme de Needleman & Wunsch et celui de Smith & Waterman. On pourra consulter la thèse de J.P. Comet [C, 1999] pour la présentation d'autres algorithmes couramment utilisés.

1.4.1 Algorithme de Needleman & Wunsch

L'algorithme de Needleman & Wunsch ([N-W]) recherche un alignement global optimal. On construit à cet effet une matrice $NW = (NW_{ij})_{0\leqslant i\leqslant m, 0\leqslant j\leqslant n}$; le terme générique NW_{ij} pour $i, j \geqslant 1$ correspond au score d'alignement optimal des séquences préfixes $a_1, \ldots, a_i$ et $b_1 \ldots, b_j$. On initialise la matrice par $NW_{00} = 0$, $NW_{i0} = i\beta, 1 \leqslant i \leqslant m$ (score linéaire d'un trou de longueur i), $NW_{0j} = j\beta, 1 \leqslant j \leqslant n$, puis on complète la matrice par récurrence en remarquant que l'alignement de $a_1, \ldots, a_i$ et $b_1 \ldots, b_j$ peut se terminer soit par $\binom{a_i}{b_j}$, soit par $\binom{a_i}{-}$, soit par $\binom{-}{b_j}$. On obtient ainsi pour $1 \leqslant i \leqslant m$ et $1 \leqslant j \leqslant n$ la relation

$$NW_{ij} = \max(NW_{i-1,j-1} + s(a_i, b_j), NW_{i-1,j} + \beta, NW_{i,j-1} + \beta).$$

Le score global des deux séquences est alors donné par NW_{mn}.

1.4.2 Algorithme de Smith & Waterman

L'algorithme de Smith & Waterman ([S-W]) recherche des alignements locaux. Il est basé sur le même principe de récurrence que celui de Needleman & Wunsch à la différence près que l'alignement local peut débuter à tout moment, ce qui revient à redémarrer un nouvel alignement juste après l'obtention d'un score négatif. En d'autres termes, la matrice correspondante SW est définie par $SW_{i0} = SW_{0j} =$ pour $0 \leqslant i \leqslant m, 0 \leqslant j \leqslant n$ et par la récurrence pour $1 \leqslant i \leqslant m$ et $1 \leqslant j \leqslant n$,

$$SW_{ij} = \max(0, NW_{i-1,j-1} + s(a_i, b_j), NW_{i-1,j} + \beta, NW_{i,j-1} + \beta).$$

2 Modèle stochastique

La séquence $\boldsymbol{A}$ est une suite de v.a. $(A_n)_{n\geqslant 1}$ à valeurs dans un alphabet $\mathcal{A}$. On introduit une fonction score $s : \mathcal{A} \longrightarrow \mathbb{R}$. Le score global est défini par $S(\boldsymbol{A}) = \sum_{k=1}^{n} s(A_k)$ et le score local par $H(\boldsymbol{A}) = \max_{1\leqslant i\leqslant j\leqslant n} \sum_{k=i}^{j} s(A_k)$. Afin de procéder a une analyse statistique de la séquence $\boldsymbol{A}$, on a besoin de déterminer la loi de probabilité des v.a. $S(\boldsymbol{A})$ et $H(\boldsymbol{A})$. En fait, en changeant de notations, on dispose d'une suite de v.a. $(X_n)_{n\geqslant 1}$ où X_n joue le rôle de $s(A_n)$ et l'on va étudier en détail la v.a. $H_n = \max_{0\leqslant i\leqslant j\leqslant n}(S_j - S_i) = \max_{0\leqslant i\leqslant j\leqslant n} \sum_{k=i}^{j} X_k$ en posant $X_0 = S_0 = 0$ et $S_n = \sum_{k=0}^{n} X_k$.

2.1 Théorèmes limites classiques dans le cas i.i.d.

On suppose les v.a. X_n, $n \geqslant 1$ indépendantes de même loi, celle d'une v.a. mère $X \in L^2$. Les trois résultats asymptotiques suivants sont classiques. On note $m = \mathbb{E}(X)$ et $\sigma^2 = \mathrm{var}\,(X)$.

Théorème 2.1 (Loi des grands nombres) *On a la convergence suivante :*

$$\frac{S_n}{n} \underset{n\to\infty}{\longrightarrow} m \text{ p.s. et en loi.}$$

Théorème 2.2 (Théorème de la limite centrale) *On la convergence suivante :*

$$\frac{S_n - nm}{\sigma\sqrt{n}} \underset{n\to\infty}{\longrightarrow} Z \text{ en loi.}$$

où Z suit la loi normale.

Théorème 2.3 (Principe de grandes déviations) *La relation limite suivante est satisfaite :*

$$\frac{1}{n}\ln\mathbb{P}(\frac{S_n}{n}-m\geqslant\varepsilon)\underset{n\to\infty}{\longrightarrow}-I(X).$$

où

$I(X)=\inf\{K(Y,X-m), Y$ v.a. absolument continue par rapport à $X-m$ telle que $\mathbb{E}(Y)=\varepsilon\}$,

K étant l'entropie de Kullback.

Selon le signe de m on a des comportements asymptotiques pour H_n différents, ce phénomène qui porte le nom de transition de phase a été décrit par Waterman, Gordon & Arratia [W-G-A, 1987], puis démontré par Arratia & Waterman [A-W, 1994] et Zhang [Z, 1995].

Théorème 2.4 (Waterman, Gordon & Arratia) *Le comportement asymptotique de H_n est donné par*

$$H_n=\begin{cases}\mathcal{O}(n) & si\ m>0,\\ \mathcal{O}(\sqrt{n}) & si\ m=0,\\ \mathcal{O}(\ln n) & si\ m<0.\end{cases}$$

De manière plus précise,

$$\begin{aligned} &si\ m>0, && \frac{H_n}{n}\underset{n\to\infty}{\overset{\text{p.s.}}{\longrightarrow}} m,\\ &si\ m=0, && \frac{H_n}{\sqrt{n}}\underset{n\to\infty}{\overset{\text{p.s.}}{\longrightarrow}} C \quad \text{pour un certain } C>0,\\ &si\ m<0, && \frac{H_n}{\ln n}\underset{n\to\infty}{\overset{\text{p.s.}}{\longrightarrow}} \frac{1}{\lambda}, \end{aligned}$$

λ étant l'unique réel positif tel que $\mathbb{E}[e^{\lambda X}]=1$.

2.2 Théorème de Karlin et al.

On considère une suite de v.a. $(X_n)_{n\geqslant 1}$ indépendantes de même loi que la v.a. mère X vérifiant les deux hypothèses $\mathbb{E}(X)<0$ et $\mathbb{P}(X>0)$ assurant une dérive vers $-\infty$ pour la suite tout en prenant des valeurs positives avec une probabilité non nulle. On introduit les instants aléatoires $\tau^{\pm}=\min\{k\geqslant 1 : S_k \gtrless 0\}$ avec la convention $\min(\emptyset)=+\infty$, puis $S^{\pm}=S_{\tau^{\pm}}$.

2.2.1 Cas continu

On suppose que X est une v.a. bornée non arithmétique, c'est-à-dire que son support n'est pas porté par un réseau. Le résultat suivant a été annoncé par Karlin & Altschul [K-A, 1990] puis démontré par Karlin, Dembo & Kawabata [K-D-K, 1990], Dembo & Karlin [D-K1, 1991] et [D-K2, 1992].

Théorème 2.5 *La v.a. H_n suit approximativement, pour n grand, la loi des extrêmes (loi de Gumbel) :*

$$\mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x) \underset{n\to\infty}{\longrightarrow} e^{-C_H e^{-\lambda x}}$$

soit encore

$$H_n - \frac{\ln n}{\lambda} \underset{n\to\infty}{\overset{\text{loi}}{\longrightarrow}} Z$$

où Z suit une loi de Gumbel, λ est l'unique réel positif tel que $\mathbb{E}[e^{\lambda X}] = 1$, et

$$C_H = \frac{\mathbb{P}(\tau^+ = \infty)\mathbb{E}[Xe^{\lambda X}]}{\lambda \mathbb{E}(S^+ e^{\lambda S^+}, \tau^+ < \infty)} = \frac{[1 - \mathbb{E}(e^{\lambda S^-})]^2}{\lambda [\mathbb{E}(\tau^-)]^2 \mathbb{E}[Xe^{\lambda X}]}.$$

De plus, si L_{H_n} désigne la longueur du premier segment qui réalise le score maximal, on a

$$\frac{L_{H_n}}{\ln n} \underset{n\to\infty}{\overset{\text{p.s.}}{\longrightarrow}} \frac{1}{\lambda \mathbb{E}(Xe^{\lambda X})}.$$

Ce résultat, qui relève de la théorie sophistiquée du renouvellement, est aujourd'hui implémenté dans les célèbres logiciels d'alignement BLAST et FASTA et utilisé par les bioinformaticiens pour analyser la composition de la séquence étudiée. Une légère adaptation du résultat de Karlin & Altschul est également exploitée pour tester la pertinence de l'alignement de deux ou plusieurs séquences.

Plan de la démonstration. On démarre par une estimation exponentielle de la queue de la distribution de $M = \max_{j\geqslant 0} S_j$ (lemme 2.8) qui repose sur la théorie du renouvellement, puis de celle de $M_0 = \max_{0\leqslant j<\tau^-} S_j$ (lemme 2.9) de laquelle on obtient un résultat asymptotique sur la loi H_{T_n} (lemme 2.10), l'instant aléatoire T_n étant celui du n-ième record négatif de la suite $(S_n)_{n\geqslant 0}$ (défini dans la section 2.3 suivante), qui conduit enfin au résultat annoncé. Auparavant, on aura besoin du résultat important suivant.

Lemme 2.6 (Factorisation de Wiener-Hopf) *Les deux factorisations suivantes sont valides pour tout $u \in]-1, 1[$ et tout $y \in \mathbb{R}$:*

$$\frac{1}{1 - u\mathbb{E}(e^{iyX})} = \Phi^-(iy)\Phi^+(iy) = \Psi^-(iy)\Psi^+(iy)$$

avec

$$\Phi^+(z) = \exp\left[\sum_{n=1}^{\infty} \frac{u^n}{n} \mathbb{E}(e^{zS_n}, S_n < 0)\right] \quad \textit{pour } \Re(z) \geqslant 0,$$

$$\Phi^-(z) = \exp\left[\sum_{n=1}^{\infty} \frac{u^n}{n} \mathbb{E}(e^{zS_n}, S_n \geqslant 0)\right] \quad \textit{pour } \Re(z) \leqslant 0,$$

$$\Psi^+(z) = \mathbb{E}\left[\sum_{n=0}^{\tau^+-1} u^n e^{zS_n}\right] \quad \textit{pour } \Re(z) \geqslant 0,$$

$$\Psi^-(z) = \frac{1}{1 - \mathbb{E}(u^{\tau^+} e^{zS^+}, \tau^+ < \infty)} \quad \textit{pour } \Re(z) \leqslant 0.$$

Les fonctions Φ^+, Φ^- (resp. Ψ^+, Ψ^-) sont analytiques sur $\{z \in \mathbb{C} : \Re(z) > 0\}$ (resp. $\{z \in \mathbb{C} : \Re(z) < 0\}$), continues et bornées ainsi que leur inverse sur $\{z \in \mathbb{C} : \Re(z) \geqslant 0\}$ (resp. $\{z \in \mathbb{C} : \Re(z) \leqslant 0\}$).

Preuve. 1. En s'appuyant sur l'identité élémentaire

$$\frac{1}{1-z} = \exp[-\ln(1-z)] = \exp\Big(\sum_{n=1}^{\infty} \frac{z^n}{n}\Big) \quad \text{pour } |z|<1,$$

on trouve

$$\begin{aligned}\frac{1}{1-u\mathbb{E}(e^{iyX})} &= \exp\Big[\sum_{n=1}^{\infty} \frac{u^n}{n}\,[\mathbb{E}(e^{iyX})]^n\Big] \\ &= \exp\Big[\sum_{n=1}^{\infty} \frac{u^n}{n}\,\mathbb{E}(e^{iyS_n})\Big] \\ &= \Phi^-(iy)\Phi^+(iy),\end{aligned}$$

la deuxième égalité provenant du fait que

$$\big[\mathbb{E}(e^{iyX})\big]^n = \mathbb{E}(e^{iyX_1}) \times \cdots \times \mathbb{E}(e^{iyX_n}) = \mathbb{E}(e^{iy(X_1+\cdots+X_n)})$$

et la troisième de

$$\mathbb{E}(e^{iyS_n}) = \mathbb{E}(e^{iyS_n}, S_n < 0) + \mathbb{E}(e^{iyS_n}, S_n \geqslant 0).$$

2. D'autre part, en utilisant

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n \quad \text{pour } |z|<1,$$

on obtient

$$\begin{aligned}\frac{1}{1-u\mathbb{E}(e^{iyX})} &= \sum_{n=0}^{\infty} u^n \big[\mathbb{E}(e^{iyX})b\big]^n \\ &= \mathbb{E}\Big[\sum_{n=0}^{\infty} u^n \mathbb{E}(e^{iyS_n})\Big] \\ &= \mathbb{E}\Big[\sum_{n=0}^{\tau-1} u^n \mathbb{E}(e^{iyS_n})\Big] + \mathbb{E}\Big[\sum_{n=\tau}^{\infty} u^n \mathbb{E}(e^{iyS_n})\Big] \\ &= \mathbb{E}\Big[\sum_{n=0}^{\tau-1} u^n \mathbb{E}(e^{iyS_n})\Big] + \mathbb{E}\Big[u^\tau \mathbb{E}(e^{iyS_\tau})\mathbb{1}_{\{\tau<\infty\}} \sum_{m=0}^{\infty} u^m e^{iy(S_{m+\tau}-S_\tau)}\Big]\end{aligned}$$

pour tout temps aléatoire τ. Si τ est un temps d'arrêt (temps non anticipatif), alors la suite $(S_{m+\tau} - S_\tau)_{m\geqslant 0}$ est une réplique de $(S_m)_{m\geqslant 0}$ indépendante de cette dernière. Ainsi

$$\begin{aligned}\frac{1}{1-u\mathbb{E}(e^{iyX})} &= \mathbb{E}\Big[\sum_{n=0}^{\tau-1} u^n \mathbb{E}(e^{iyS_n})\Big] + \mathbb{E}[u^\tau \mathbb{E}(e^{iyS_\tau}), \tau<\infty]\,\mathbb{E}\Big[\sum_{m=0}^{\infty} u^m e^{iyS_m}\Big] \\ &= \mathbb{E}\Big[\sum_{n=0}^{\tau-1} u^n \mathbb{E}(e^{iyS_n})\Big] + \mathbb{E}[u^\tau \mathbb{E}(e^{iyS_\tau}), \tau<\infty]\,\frac{1}{1-u\mathbb{E}(e^{iyX})}\end{aligned}$$

ce qui entraîne finalement

$$\frac{1}{1-u\mathbb{E}(e^{iyX})} = \frac{\mathbb{E}\big[\sum_{n=0}^{\tau-1} u^n \mathbb{E}(e^{iyS_n})\big]}{1-\mathbb{E}[u^\tau \mathbb{E}(e^{iyS_\tau}), \tau<\infty]}.$$

□

Remarque. On peut démontrer en utilisant l'identité en loi de $(S_n - S_{n-k})_{1\leqslant k\leqslant n}$ et $(S_k)_{1\leqslant k\leqslant n}$ que

$$\Psi^+(z) = \frac{1}{1-\mathbb{E}(u^{\tau^-}e^{zS^-})} \quad \text{pour } \Re(z)\geqslant 0,$$
$$\Psi^-(z) = \mathbb{E}\Big[\sum_{n=0}^{\tau^- -1} u^n e^{zS_n}\Big] \quad \text{pour } \Re(z)\leqslant 0.$$

Lemme 2.7 *Les relations suivantes sont satisfaites :*

$$\mathbb{E}(\tau^-) = \frac{1}{\mathbb{P}(\tau^+=\infty)},$$
$$\mathbb{E}(e^{\lambda S^+}, \tau^+<\infty) = 1,$$
$$\mathbb{E}(S^+e^{\lambda S^+}, \tau^+<\infty) = \frac{E(Xe^{\lambda X})}{1-\mathbb{E}(e^{\lambda S^-})}.$$

Preuve. Il est aisé de voir que la fonction définie sur $\mathbb{C}$ par

$$f(z) = \begin{cases} \dfrac{\Psi^+(z)}{\Phi^+(z)} & \text{si } \Re(z)\geqslant 0, \\ \dfrac{\Phi^-(z)}{\Psi^-(z)} & \text{si } \Re(z)\leqslant 0, \end{cases}$$

est une fonction entière borné, donc constante et que cette constante vaut 1. Ceci assure l'unicité de la factorisation de Wiener-Hopf de $\dfrac{1}{1-u\mathbb{E}(e^{iyX})}$:

$$\Phi^+ = \Psi^+ \quad \text{et} \quad \Phi^- = \Psi^-.$$

On peut alors prouver les relations annoncées à partir de ces identités. □

Lemme 2.8 *On a l'équivalence asymptotique suivante :*

$$\mathbb{P}(M>x) \underset{x\to+\infty}{\sim} C_M e^{-\lambda x}$$

où $C_M = \dfrac{\mathbb{P}(\tau^+=\infty)}{\lambda\mathbb{E}(S^+e^{\lambda S^+}, \tau^+<\infty)}$.

Preuve. On montre facilement que la fonction de répartition F_M de M, $F_M(x) = \mathbb{P}(M\leqslant x)$, satisfait à l'équation intégrale

$$F_M(x) = 1 - F_{\tau^+}(+\infty) + \int_0^x F_M(x-y)\, dF_{S^+}(y).$$

Comme $F_{S^+}(+\infty) = \mathbb{P}(S^+<\infty) = \mathbb{P}(\tau^+<\infty) < 1$, la loi de S^+ est impropre (masse à l'infini). Par contre, la relation

$$\int_0^\infty e^{\lambda y}\, dF_{S^+}(y) = \mathbb{E}(e^{\lambda S^+}, \tau^+<\infty) = 1$$

(cf. lemme 2.7) comble cette lacune, ce qui conduit à introduire la fonction $\varphi(x) = [1-F_M(x)]\, e^{\lambda x}$, laquelle vérifie l'équation du renouvellement

$$\varphi(x) = \psi(x) + \int_0^x \varphi(x-y)\, e^{\lambda y}\, dF_{S^+}(y)$$

où l'on a posé $\psi(x) = [F_{S^+}(+\infty) - F_{S^+}(x)]\, e^{\lambda x}$. Le théorème du renouvellement appliqué à la mesure de probabilité non arithmétique $e^{\lambda y}\, dF_{S^+}(y)$ fournit la limite de φ en $+\infty$:

$$\lim_{x\to+\infty} \varphi(x) = \frac{\int_0^{+\infty} \psi(x)\, dx}{\int_0^{+\infty} xe^{\lambda x}\, dF_{S^+}(x)}.$$

D'une part,

$$\int_0^{+\infty} \psi(x)\, dx = \frac{1}{\lambda}\int_0^{+\infty} (e^{\lambda x} - 1)\, dF_{S^+}(x) = \frac{1}{\lambda}\,\mathbb{E}(e^{\lambda S^+} - 1, \tau^+ < \infty) = \frac{1}{\lambda}\,\mathbb{P}(\tau^+ = +\infty)$$

et d'autre part,

$$\int_0^{+\infty} xe^{\lambda x}\, dF_{S^+}(x) = \mathbb{E}(S^+ e^{\lambda S^+}, \tau^+ < \infty)$$

puis le résultat annoncé s'en déduit aussitôt. □

Lemme 2.9 *On a l'équivalence asymptotique suivante :*

$$\mathbb{P}(M_0 > x) \underset{x\to+\infty}{\sim} C_{M_0} e^{-\lambda x}$$

où $C_{M_0} = C_M[1 - \mathbb{E}(e^{\lambda S^-})]$.

Preuve. On montre aisément la relation suivante

$$\mathbb{P}(M > x) = \mathbb{P}(M_0 > x) + \int_{-\infty}^{0} \mathbb{P}(M > x - y)\, \mathbb{P}(S^- \in dy, M_0 \leqslant x)$$

de laquelle on déduit le lemme 2.9 grâce au lemme 2.8. □

Lemme 2.10 *La limite suivante est valide pour tout réel* x *:*

$$\mathbb{P}(H_{T_n} \leqslant \frac{\ln n}{\lambda} + x) \underset{n\to\infty}{\longrightarrow} e^{-C_{M_0} e^{-\lambda x}}.$$

Preuve. On verra dans la section 2.3 que $H_{T_n} = \max_{0\leqslant k\leqslant n-1} M_k$, les v.a. M_k étant indépendantes de même loi que M_0. Il est alors immédiat que

$$\mathbb{P}(H_{T_n} \leqslant y) = \mathbb{P}(M_0 \leqslant y)^n.$$

Ainsi, lorsque $n \to +\infty$,

$$\begin{aligned}
\mathbb{P}(H_{T_n} \leqslant \frac{\ln n}{\lambda} + x) &= \big[1 - (C_{M_0} + o(1))e^{-\lambda(x + \frac{\ln n}{\lambda})}\big]^n \\
&= \big[1 - \frac{1}{n}(C_{M_0} + o(1))e^{-\lambda x}\big]^n \\
&\sim e^{-C_{M_0} e^{-\lambda x}}.
\end{aligned}$$

□

Fin de la démonstration du théorème. Par monotonie, il est clair que l'encadrement suivant est satisfait :

$$\mathbb{P}(H_{T_{N(n)+1}} \leqslant y) \leqslant \mathbb{P}(H_n \leqslant y) \leqslant \mathbb{P}(H_{T_{N(n)}} \leqslant y),$$

la v.a. $N(n)$ étant le nombre de records négatifs de la suite $(S_n)_{n\geqslant 0}$ défini dans la section 2.3 suivante. D'autre part, la loi des grands nombres assure que $\frac{T_m}{m} \underset{m\to\infty}{\longrightarrow} \mathbb{E}(T_1) = \mathbb{E}(\tau^-)$ p.s., ce qui implique, en choisissant $m = N(n)$,

$$\frac{N(n)}{n} \underset{n\to\infty}{\longrightarrow} \mu = \frac{1}{\mathbb{E}(\tau^-)} \text{ p.s.}$$

De ces faits, on tire, pour tout $\varepsilon > 0$,

$$\mathbb{P}(H_{T_{[n(\mu+\varepsilon)]+1}} \leqslant \frac{\ln n}{\lambda} + x) + o(1) \leqslant \mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x) \leqslant \mathbb{P}(H_{T_{[n(\mu-\varepsilon)]}} \leqslant \frac{\ln n}{\lambda} + x) + o(1)$$

puis en appliquant le lemme 2.10,

$$\lim_{n\to+\infty} \mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x) = \lim_{m\to+\infty} \mathbb{P}(H_{T_m} \leqslant \frac{\ln m}{\lambda} - \frac{\ln \mu}{\lambda} + x) = e^{-\mu C_{M_0} e^{-\lambda x}},$$

ce qui termine la preuve du théorème. □

2.2.2 Cas discret

Énonçons le résultat dans le cas où X est une v.a. arithmétique de pas δ, *i.e.* X est à valeurs dans le réseau $\delta\mathbb{Z}$ et bornée.

Théorème 2.11 *Le comportement asymptotique de la v.a. H_n est décrit par les inégalités suivantes :*

$$e^{-C_H^+ e^{-\lambda x}} = \liminf_{n\to+\infty} \mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x) < \limsup_{n\to+\infty} \mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x) = e^{-C_H^- e^{-\lambda x}}$$

où

$$C_H^+ = C_H^- e^{\delta\lambda} \quad \text{et} \quad C_H^- = \frac{\delta\lambda}{e^{\delta\lambda} - 1} C_H.$$

La démarche est la même que dans le cas continu, elle repose sur une équation du renouvellement discrète. La différence essentielle avec le cas continu est que la quantité $\mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x)$ n'admet pas de limite lorsque $n \to +\infty$. Karlin & Altschul ([K-A]) utilisent l'approximation

$$\mathbb{P}(H_n \leqslant \frac{\ln n}{\lambda} + x) \approx \Big(1 - \frac{C_{M_0} e^{-\lambda x}}{n}\Big)^{\frac{n}{\mathbb{E}(\tau^-)}+1}.$$

Exemples.

1. X prend ses valeurs dans $\{-1, 0, 1, \ldots, m\}$. On peut voir facilement que $S^- = -1$ et donc

$$C_H = (1 - e^{-\lambda})^2 \frac{[E(X)]^2}{\lambda\, \mathbb{E}(X e^{\lambda X})}.$$

2. X prend ses valeurs dans $\{-m, \ldots, -1, 0, 1\}$. Dans ce cas, on a $S^+ = 0$ sur $\{X_1 \leqslant 0, \tau^+ < \infty\}$ et $S^+ = 1$ sur $\{X_1 = 1\}$. On obtient

$$C_H = \frac{1}{\lambda}(1 - e^{-\lambda})^2\, \mathbb{E}(X e^{\lambda X}).$$

2.3 Cas d'une suite de v.a. i.i.d. à valeurs entières

Soit $(X_n)_{n\geqslant 1}$ une suite de v.a. i.i.d. pour l'instant à valeurs dans $\mathbb{R}$. On introduit la suite des instants de records négatifs

$$T_0 = 0,\ \ T_{n+1} = \min\{k > T_n : S_k - S_{T_n} < 0\},$$

ainsi que le nombre de records dans la suite finie $S_0, \ldots, S_n$:

$$N(n) = \max\{k \geqslant 0 : T_k \leqslant n\};$$

$T_{N(n)}$ est l'instant du dernier record inférieur ou égal à n et l'on a

$$T_{N(n)} \leqslant n < T_{N(n)+1} \quad \text{et} \quad N(T_{N(n)}) = n.$$

On définit alors le processus d'excursion

$$U_0 = 0,\ \ U_n = S_n - S_{T_{N(n)}}, n \geqslant 1.$$

On pose enfin

$$M_k = \max_{T_k \leqslant j < T_{k+1}} (S_j - S_{T_k}).$$

On a en particulier $M_0 = \max_{0\leqslant j<\tau^-} S_j$.

Proposition 2.12 *On a pour tout $n \geqslant 0$*

$$\begin{aligned} U_{T_n} &= 0, \\ U_{n+1} &= (U_n + X_{n+1})^+, \\ H_n &= \max_{0\leqslant k\leqslant n} U_k. \end{aligned}$$

Preuve. 1. On a $U_{T_n} = S_{T_n} - S_{T_{N(T_n)}} = 0$ puisque $N(T_n) = n$.
2. $T_{N(n)}$ étant le dernier record négatif avant n, on a $S_{T_n} \leqslant S_n$ et alors $U_n \geqslant 0$.
3. On a $U_{n+1} = S_{n+1} - S_{T_{N(n+1)}} = S_n + X_{n+1} - S_{T_{N(n+1)}}$. Deux cas sont à distinguer.

(a) Soit $T_{N(n+1)} = n+1$ et alors $U_{n+1} = U_{T_{N(n+1)}} = 0$, puis

$$U_n + X_{n+1} = S_n - S_{T_{N(n)}} + X_{n+1} = S_{n+1} - S_{T_{N(n)}} = S_{T_{N(n+1)}} - S_{T_{N(n)}} < 0$$

par définition de la suite $T_{N(n)}$. D'où $(U_n + X_{n+1})^+ = 0 = U_{n+1}$.

(b) Soit $T_{N(n+1)} \leqslant n$ et alors $T_{N(n+1)} = T_{N(n)}$, puis

$$U_{n+1} = S_n - S_{T_{N(n)}} + X_{n+1} = U_n + X_{n+1}.$$

Comme $U_{n+1} \geqslant 0$, on a $U_n + X_{n+1} \geqslant 0$ ce qui entraîne $(U_n + X_{n+1})^+ = U_{n+1}$.

4. Remarquons que

$$H_n = \max_{0\leqslant i\leqslant j\leqslant n} (S_j - S_i) = \max_{0\leqslant j\leqslant n} [\max_{0\leqslant i\leqslant j} (S_j - S_i)].$$

Or, pour tout $i \leqslant j$, $S_{T_{N(j)}} \leqslant S_i$, donc $S_j - S_i \leqslant S_j - S_{T_{N(j)}} = U_j$ et pour $i = T_{N(j)}$, $S_j - S_i = U_j$. Ainsi $\max_{0\leqslant i\leqslant j}(S_j - S_i) = U_j$ ce qui prouve que $H_n = \max_{0\leqslant j\leqslant n} U_j$. □

Corollaire 2.13 *1. La loi de U_n est identique à celle de $\max_{0\leqslant k\leqslant n} S_k$.*
2. On a la relation

$$H_{T_n} = \max_{0\leqslant k\leqslant n-1} M_k.$$

Preuve. 1. En itérant la formule $U_n = \max(U_{n-1} + X_n, 0)$, on trouve

$$U_n = \max(X_1 + \cdots + X_n, X_2 + \cdots + X_n, \ldots, X_{n-1} + X_n, X_n, 0),$$

ce qui s'écrit en effectuant le retournement de temps $k \in \{1, \ldots, n\} \mapsto n - k + 1 \in \{1, \ldots, n\}$, *i.e.* en posant $X'_k = X_{n-k+1}$ et $S'_k = \sum_{i=1}^{k} X'_i$,

$$U_n = \max(X'_1 + \cdots + X'_n, X'_1 + \cdots + X'_{n-1}, \ldots, X'_1 + X'_2, X'_1, 0) = \max_{0 \leqslant k \leqslant n} S'_k.$$

La suite $(X_1, \ldots, X_n)$ a même loi que $(X'_1, \ldots, X'_n)$, ce qui prouve l'identité en loi des v.a. U_n et $\max_{0 \leqslant k \leqslant n} S_k$.
2. Puisque $U_{T_n} = 0$, on a

$$H_{T_n} = \max_{0 \leqslant j \leqslant T_n - 1} U_j = \max_{0 \leqslant k \leqslant n-1} \Big(\max_{T_k \leqslant j < T_{k+1}} U_j \Big).$$

Or pour tout $j \in [T_k, T_{k+1}[$, $N(j) = k$, donc $U_j = S_j - S_{T_k}$ et alors

$$\max_{T_k \leqslant j < T_{k+1}} U_j = M_k.$$

□

On suppose maintenant la v.a. X à valeurs dans $\mathbb{Z}$ de loi déterminée par les nombres $p_k = \mathbb{P}(X = k)$, $k \in \mathbb{Z}$. Posons $F_k = \mathbb{P}(X \leqslant k)$. Soit $a \in \mathbb{N}$, $\tau_a = \min\{n \geqslant 1 : U_n \geqslant a\}$ et U^a le processus absorbé en a

$$U^a_n = \begin{cases} U_n & \text{si } n < \tau_a, \\ a & \text{si } n \geqslant \tau_a. \end{cases}$$

Proposition 2.14 (Daudin & Mercier [D-M, 1999]) *1. La suite $(U_n)_{n \geqslant 0}$ est une chaîne de Markov sur l'espace d'états $\mathbb{N}$ de matrice de transition*

$$\boldsymbol{P} = \begin{pmatrix} F_0 & p_1 & p_2 & \\ F_{-1} & p_0 & p_1 & \ddots \\ F_{-2} & p_{-1} & p_0 & \ddots \\ \vdots & & \ddots & \ddots \end{pmatrix}.$$

2. La suite $(U^a_n)_{n \geqslant 0}$ est une chaîne de Markov sur l'espace d'états $\{0, 1, \ldots, a\}$ de matrice de transition

$$\boldsymbol{P}^a = \left(\begin{array}{c|cccccc} F_0 & p_1 & \cdots & p_j & \cdots & p_{a-1} & 1 - F_{a-1} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ F_{-i} & p_{1-i} & \cdots & p_{j-i} & \cdots & p_{a-i} & 1 - F_{a-i} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ F_{1-a} & p_{2-a} & \cdots & p_{j-(a-1)} & \cdots & p_0 & 1 - F_0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right).$$

Preuve. 1. Introduisons la famille des probabilités de transition

$$p_{ij} = \mathbb{P}(U_{n+1} = j | U_n = i), \ i, j \in \mathbb{N}.$$

On a

$$p_{ij} = \mathbb{P}((X_{n+1} + i)^+ = j) = \begin{cases} \mathbb{P}(X = j - i) = p_{j-i} & \text{si } j \geqslant 1, \\ \mathbb{P}(X \leqslant -i) = F_{-i} & \text{si } j = 0. \end{cases}$$

D'où la matrice $\boldsymbol{P}$.
2. Introduisons de même

$$p^a_{ij} = \mathbb{P}(U^a_{n+1} = j | U^a_n = i), \ i, j \in \{0, 1, \ldots, a\}.$$

On a

- pour $i = a$: $p^a_{aa} = 1$ et $p^a_{aj} = 0$ pour $j \in \{0, \ldots, a-1\}$, il y a absorption en a.
- pour $i \in \{0, 1, \ldots, a-1\}$:
  - si $j = 0$, $p^a_{i0} = \mathbb{P}(U_{n+1} = 0 | U_n = i) = \mathbb{P}(X \leqslant -i) = F_{-i}$;
  - si $j = a$, $p^a_{ia} = \mathbb{P}(U_{n+1} \geqslant a | U_n = i) = \mathbb{P}(X \geqslant a - i) = 1 - F_{a-i-1}$;
  - si $j \in \{1, \ldots, a-1\}$, $p^a_{ij} = p_{ij} = p_{j-i}$.

□

On peut alors exprimer la loi de H_n en fonction de la matrice $\boldsymbol{P}^a$ selon

$$\mathbb{P}(H_n < a) = \mathbb{P}(\tau_a > n) = 1 - \mathbb{P}(U^a_n = a) = 1 - (1\,0\,\ldots\,0)(\boldsymbol{P}^a)^n \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

2.4 Cas d'une suite de v.a. i.i.d. à valeurs dans $\{-b_s, \ldots, -b_1, a_1, \ldots, a_r\}$

On suppose que X est à valeurs dans l'alphabet $\{-b_s, \ldots, -b_1, a_1, \ldots, a_r\}$ où $a_1, \ldots, a_r, b_1, \ldots, b_s$ sont des entiers naturels premiers entre eux vérifiant $-b_s \leqslant \ldots \leqslant -b_1 < 0 \leqslant a_1 \leqslant \ldots \leqslant a_r$. La loi de X est donnée par les nombres $p_i = \mathbb{P}(X = a_i)$, $1 \leqslant i \leqslant r$ et $q_j = \mathbb{P}(X = -b_j)$, $1 \leqslant j \leqslant s$ avec $\sum_{i=1}^r p_i + \sum_{j=1}^s q_j = 1$.

Proposition 2.15 (Mercier [M, 1999]) *La chaîne de Markov $(U_n)_{n \geqslant 0}$ est irréductible et admet la loi de M comme loi invariante.*

Preuve. 1. On prouve l'irréductibilité de la chaîne $(U_n)_{n \geqslant 0}$ en montrant que tous les états $x \in \mathbb{N}$ communiquent avec l'état 0.

–

(a) Il existe un chemin allant de x vers 0 de probabilité positive obtenu en effectuant $n_1 = [x/b_s] + 1$ pas de longueur $-b_s$ (on a $x - (n_1 - 1)b_s \geqslant 0 > x - n_1 b_s$) :

$$\begin{aligned} \mathbb{P}_x(U_{n_1} = 0) &\geqslant \mathbb{P}(U_1 = x - b_s, \ldots, U_{n_1 - 1} = x - (n_1 - 1)b_s, U_{n_1} = 0 | U_0 = x) \\ &\geqslant \mathbb{P}(X_1 = -b_s, \ldots, X_{n_1} = -b_s) = q_s^{n_1} > 0. \end{aligned}$$

(b) On va construire un chemin allant de 0 vers x de probabilité positive en utilisant le fait que les entiers $a_1, \ldots, a_r, b_1, \ldots, b_s$ sont premiers entre eux.

Lemme 2.16 *La décomposition suivante est satisfaite :*

$$\sum_{i=1}^r a_i \mathbb{N} + \sum_{j=1}^s (-b_j)\mathbb{N} = \mathbb{Z}.$$

D'après le lemme, l'entier naturel x peut s'écrire sous la forme $\sum_{i=1}^r \lambda_i a_i + \sum_{j=1}^s \mu_j(-b_j)$, ce qui fournit un chemin de 0 vers x obtenu en effectuant λ_1 pas de longueur $a_1, \ldots, \lambda_r$ pas de longueur a_r, μ_1 pas de longueur $-b_1, \ldots, \mu_s$ pas de longueur $-b_s$, ce qui fait un total de $\lambda_1 + \cdots + \mu_s$ pas tout en restant à valeurs positives. La probabilité d'un tel chemin minore $\mathbb{P}_0(U_{\lambda_1 + \cdots + \mu_s} = x)$:

$$\begin{aligned}\mathbb{P}_0(U_{\lambda_1+\cdots+\mu_s}=x) \geqslant \mathbb{P}(&X_1=\ldots=X_{\lambda_1}=a_1,\\ &X_{\lambda_1+1}=\ldots=X_{\lambda_1+\lambda_2}=a_2,\\ &\cdots\\ &X_{\lambda_1+\cdots+\lambda_{r-1}+1}=\ldots=X_{\lambda_1+\cdots+\lambda_r}=a_r,\\ &X_{\lambda_1+\cdots+\lambda_r+1}=\ldots=X_{\lambda_1+\cdots+\lambda_r+\mu_1}=-b_1,\\ &X_{\lambda_1+\cdots+\mu_1+1}=\ldots=X_{\lambda_1+\cdots+\mu_2}=-b_2,\\ &\cdots\\ &X_{\lambda_1+\cdots+\mu_{s-1}+1}=\ldots=X_{\lambda_1+\cdots+\mu_s}=-b_s)\end{aligned}$$

$$\geqslant \prod_{i=1}^{r} p_i^{\lambda_i} \prod_{j=1}^{s} q_j^{\mu_j} > 0.$$

2. On introduit de nouvelles notations : soit $(X_{-n})_{n\geqslant 1}$ une suite de v.a. i.i.d. de même loi que X et indépendante de la suite $(X_n)_{n\geqslant 1}$. On adoptera la convention $\sum_{n=a}^{b}=0$ si $a>b$. Pour montrer que la loi de M est invariante, on suppose que U_0 suit cette loi, ce qui revient à supposer que U_0 peut s'écrire sous la forme

$$U_0=\sup_{n\geqslant 0}\sum_{i=0}^{n} X_{-i}=\sup_{n\geqslant -1}\sum_{i=-n}^{0} X_i.$$

La relation de récurrence $U_{n+1}=(U_n+X_{n+1})^+$ conduit simplement à la relation $U_m=\sup_{n\geqslant -m-1}\sum_{i=-n}^{m} X_i$ de laquelle on déduit l'identité en loi

$$(U_m,U_{m+1},\ldots,U_{m+k}) \overset{\text{loi}}{=} (U_0,\ldots,U_k),$$

prouvant ainsi que la chaîne $(U_n)_{n\geqslant 0}$ est stationnaire dès que U_0 suit la loi de M. La loi de M est donc bien une loi invariante. □

Preuve du lemme. Soit $d^+=\text{pgcd}(a_1,\ldots,a_r)$ et $d^-=\text{pgcd}(b_1,\ldots,b_s)$. On a $d^+\mathbb{Z}=\sum_{i=1}^{r} a_i\mathbb{Z}$ et $d^-\mathbb{Z}=\sum_{j=1}^{s} b_j\mathbb{Z}$. Des considérations d'arithmétique élémentaire permettent de voir que

$$\sum_{i=1}^{r} a_i\mathbb{N}\supset d^+\mathbb{N}\cap[N^+,+\infty[\quad\text{et}\quad \sum_{j=1}^{s} b_j\mathbb{N}\supset d^-\mathbb{N}\cap[N^-,+\infty[$$

pour des entiers naturels N^+ et N^-. On a alors

$$\begin{aligned}\sum_{i=1}^{r} a_i\mathbb{N}+\sum_{j=1}^{s}(-b_j)\mathbb{N} &\supset \{ad^+-bd^-, a,b\in\mathbb{N}, a\geqslant N^+, b\geqslant N^-\}\\ &= N^+d^+-N^-d^-+\{ad^+-bd^-, a,b\in\mathbb{N}\}.\end{aligned}$$

D'autre part, les entiers positifs d^+ et d^- sont premiers entre eux, il existe donc $\alpha,\beta\in\mathbb{Z}$ tels que $\alpha d^+-\beta d^-=1$. En fait, il existe plus précisément $\alpha_0,\beta_0\in\mathbb{N}$ et $\alpha_1,\beta_1\in-\mathbb{N}$ tels que $\alpha_0 d^+-\beta_0 d^-=\alpha_1 d^+-\beta_1 d^-=1$. Soit maintenant $n\in\mathbb{Z}$. On a les décompositions

$$n=(\alpha_0 n)d^+-(\beta_0 n)d^-=(\alpha_1 n)d^+-(\beta_1 n)d^-$$

avec $\alpha_0 n,\beta_0 n\in\mathbb{N}$ si $n\in\mathbb{N}$ et $\alpha_1 n,\beta_1 n\in\mathbb{N}$ si $n\in-\mathbb{N}$. Ainsi n est de la forme ad^+-bd^- avec $a,b\in\mathbb{N}$. Ceci prouve bien que $\sum_{i=1}^{r} a_i\mathbb{N}+\sum_{j=1}^{s}(-b_j)\mathbb{N}=\mathbb{Z}$. □

Un résultat classique de la théorie des chaînes de Markov stipule que si une chaîne irréductible admet une loi invariante, alors la chaîne est récurrente et admet une unique loi invariante. De plus, l'état 0 est apériodique puisque $p_{00}=F_0>0$; la chaîne de Markov $(U_n)_{n\geqslant 0}$ irréductible est alors apériodique. En conséquence, $(U_n)_{n\geqslant 0}$ est ergodique de loi limite celle de M.

On peut décrire plus explicitement la loi de M. Notons $\pi_j = \mathbb{P}(M = j)$, $j \in \mathbb{N}$ et $\boldsymbol{\pi} = (\pi_j)_{j \geqslant 0}$ le vecteur-ligne de la loi de M. L'invariance se traduit par la relation $\boldsymbol{\pi} = \boldsymbol{\pi P}$, soit :

$$\pi_0 = \sum_{i=0}^{b_1} \pi_i \sum_{k=1}^{s} q_k + \sum_{j=2}^{s} \sum_{i=b_{j-1}+1}^{b_j} \pi_i \sum_{k=j}^{s} q_k$$

$$\pi_j = \begin{cases} \displaystyle\sum_{k=1}^{s} q_k \pi_{j+b_k} & \text{si } 1 \leqslant j < a_1, \\ \displaystyle\sum_{k=1}^{s} q_k \pi_{j+b_k} + \sum_{k=1}^{i} p_k \pi_{j-a_k} & \text{si } a_i \leqslant j < a_{i+1} \text{ pour } 1 \leqslant i \leqslant r-1, \\ \displaystyle\sum_{k=1}^{s} q_k \pi_{j+b_k} + \sum_{k=1}^{r} p_k \pi_{j-a_k} & \text{si } j \geqslant a_r \end{cases}$$

avec de plus $\sum_{j=0}^{\infty} \pi_j = 1$. La dernière égalité (pour $j \geqslant a_r$) montre alors que la loi $\boldsymbol{\pi}$ est définie par une récurrence linéaire. L'équation caractéristique associée est

$$1 = \sum_{k=1}^{s} q_k \rho^{b_k} + \sum_{k=1}^{r} p_k \rho^{-a_k}$$

ou encore $P(\rho) = 0$, P étant le polynôme

$$P(x) = \sum_{k=1}^{r} p_k x^{a_r - a_k} + \sum_{k=1}^{s} q_k x^{a_r + b_k} - x^{a_r} = x^{a_r}[E(x^{-X}) - 1].$$

Soit $\rho_1, \ldots, \rho_m$ les racines distinctes de P de module strictement inférieur à 1.

Théorème 2.17 (Mercier, 1999) *La loi de M est une combinaison linéaire des suites géométriques $(\rho_1^n)_{n\geqslant 0}, \ldots, (\rho_m^n)_{n\geqslant 0}$ lorsque les racines $\rho_1, \ldots, \rho_m$ sont simples et plus généralement des suites $(n^l \rho_i^n)_{n\geqslant 0}$, $0 \leqslant l \leqslant m_i - 1, 1 \leqslant i \leqslant m$ si m_i est la multiplicité de ρ_i.*

REMARQUE. On montre facilement que la condition $\mathbb{E}(X) < 0$ entraîne que le polynôme P n'admet que des deux racines réelles positives et qu'elles sont simples ; ce sont 1 et $e^{-\lambda} \in]0, 1[$. De plus, parmi toutes les racines de module strictement inférieur à 1, $e^{-\lambda}$ est celle de module maximum et ce strictement.

Exemple.

1. Recherche des segments les plus amphotères

L'alphabet $\mathcal{A}$ est celui des acides aminés :

$$\mathcal{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

défini selon le dictionnaire

acide aminé	code	acide aminé	code	acide aminé	code	acide aminé	code
alanine	A	glycine	G	méthionine	M	sérine	S
cystéine	C	histine	H	aspargine	N	thréonine	T
acide aspartique	D	isoleucine	I	proline	P	valine	V
acide glutamique	E	lysine	K	glutamine	Q	tryptophane	W
phénylalanine	F	leucine	L	arginine	R	tyrosine	Y

Les acides aminés peuvent être chargés positivement ou négativement suivant le milieu ambiant où ils se trouvent. Ils se placent en général à la surface des molécules, dans les boucles, et sont liés aux réactions chimiques. Karlin & Altschul ([K-A]) introduisent la fonction de score donnée par

$$s(a) = \begin{cases} 2 & \text{pour } a \in \{D, E, H, K, R\}, \\ -1 & \text{pour les autres acides aminés.} \end{cases}$$

Cette situation correspond au cas où $-b_1 = -1$, $a_1 = 2$, $p_1 = p_D + p_E + p_H + p_K + p_R$ et $q_1 = 1 - p$. On pose $p = p_1 = \mathbb{P}(X \geqslant 0)$ et l'on se place dans le cas où $\mathbb{E}(X) < 0$, *i.e.* $p < \frac{1}{3}$. La matrice de transition de la chaîne $(U_n)_{n \geqslant 0}$ s'écrit

$$\boldsymbol{P} = \begin{pmatrix} q & 0 & p & 0 & 0 & \cdots \\ q & 0 & 0 & p & 0 & \cdots \\ 0 & q & 0 & 0 & p & \\ \vdots & & \ddots & & & \ddots \end{pmatrix}.$$

Le polynôme P est donné par $P(x) = (1-p)x^3 - x^2 + p$ et ses racines qui sont dans $]-1, 1[$ par $\rho = \frac{p + \sqrt{p(4-3p)}}{2(1-p)}$ et $\rho' = \frac{p - \sqrt{p(4-3p)}}{2(1-p)}$. On a $\rho' < 0$, $|\rho'| < \rho < 1$. D'où la loi de M :

$$\mathbb{P}(M = n) = \frac{1 - 3p}{\sqrt{p(4-3p)}} \, (\rho^{n+1} - \rho'^{n+1}).$$

Calculons les valeurs des constantes $C_H^- = \frac{\lambda}{e^\lambda - 1} \, C_H$ et $C_H^+ = \frac{\lambda}{1 - e^{-\lambda}} \, C_H$; C_H a pour expression d'après l'un des exemples suivant le théorème 2.11

$$C_H = e^{-\lambda}(1 - e^{-\lambda}) \, \frac{[E(X)]^2}{\mathbb{E}(Xe^{\lambda X})}.$$

Rappelons également que $\rho = e^{-\lambda}$ (cf. remarque ci-dessus). On a alors

$$\begin{aligned} \mathbb{E}(X) &= 3p - 1, \\ \mathbb{E}(Xe^{\lambda X}) &= 2pe^{2\lambda} - (1-p)e^{-\lambda} = \frac{1}{\rho^2}(3p - \rho^2) \end{aligned}$$

et donc

$$\begin{aligned} C_H^- &= \frac{\rho^3(1-\rho)^2(1-3p)^2}{3p - \rho^2} = \frac{(1-3p)^2\rho^2}{\sqrt{p(4-3p)}}, \\ C_H^+ &= \frac{1}{\rho} \, C_H^- = \frac{(1-3p)^2\rho}{\sqrt{p(4-3p)}}. \end{aligned}$$

2. Recherche des segments les plus hydrophobes

On recherche des régions anormalement hydrophobes dans une protéine. La fonction de score qui a été introduite par Karlin & Altschul est donnée par

$$s(a) = \begin{cases} 2 & \text{pour } a \in \{I, L, V\}, \\ 1 & \text{pour } a \in \{A, C, F, M\}, \\ 0 & \text{pour } a \in \{G, P, S, T, W, Y\}, \\ -1 & \text{pour } a \in \{D, E, H, N, Q\}, \\ -2 & \text{pour } a \in \{K, R\}. \end{cases}$$

Cette situation correspond au cas où $-b_2 = -2$, $-b_1 = -1$, $a_1 = 0$, $a_2 = 1$, $a_3 = 2$, et

$$\begin{aligned} q_2 &= p_K + p_R, \\ q_1 &= p_D + p_E + p_H + p_N + p_Q, \\ p_1 &= p_G + p_P + p_S + p_T + p_W + p_Y, \\ p_2 &= p_A + p_C + p_F + p_M, \\ p_3 &= p_I + p_L + p_V. \end{aligned}$$

On pose $p = \mathbb{P}(X \geqslant 0) = p_1 + p_2 + p_3$ et on se place dans le cas où $\mathbb{E}(X) < 0$. La matrice de transition de la chaîne $(U_n)_{n\geqslant 0}$ s'écrit cette fois

$$\boldsymbol{P} = \begin{pmatrix} p_1+q_1+q_2 & p_2 & p_3 & 0 & 0 & 0 & \cdots \\ q_1+q_2 & p_1 & p_2 & p_3 & 0 & 0 & \cdots \\ q_2 & q_1 & p_1 & p_2 & p_3 & 0 & \cdots \\ 0 & q_2 & q_1 & p_1 & p_2 & p_3 & \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

Le polynôme P est donné par $P(x) = q_2x^4 + q_1x^3 - x^2 + p_2x + (p_1 + p_3) = (x-1)[q_2x^3 + (q_1 + q_2)x^2 + (q_1 + q_2 - 1)x - (p_1 + p_3)]$. Il admet deux racines dans $]-1, 1[$, ρ et ρ', elles sont telles que $\rho' < 0$, $|\rho'| < \rho < 1$. La loi de M a la forme suivante :

$$\mathbb{P}(M = n) = \alpha\rho^n + \beta\rho'^n.$$

RÉFÉRENCES

[A-W] R. Arratia & M.S. Waterman. A phase transition for the score in matching random sequences allowing deletions, Ann. Appl. Probab. 4 (1994), no. 1, 200–225.

[C] J.P. Comet. Programmation dynamique et alignement de séquences biologiques, Thèse de Doctorat de l'université de Compiègne, 1998.

[D-K1] A. Dembo & S. Karlin. Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of i.i.d. variables, Ann. Probab. 19 (1991), no. 4, 1737–1755.

[D-K2] A. Dembo & S. Karlin. Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of Markov variables, Ann. Probab. 19 (1991), no. 4, 1756–1767.

[D-M] J.J. Daudin & S. Mercier. Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées, C.R. Acad. Sci. Paris 329, 1 (1999), 815–820.

[K-A] S. Karlin & S.F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc. Natl. Acad. Sci. U.S.A. 87 (1990), 2264–2268.

[K-D] S. Karlin & A. Dembo. Limit distributions of maximal segmental score among Markov-dependent partial sums, Adv. in Appl. Probab. 24 (1992), no. 1, 113–140.

[K-D-K] S. Karlin, A. Dembo & T. Kawabata. Statistical composition of high-scoring segments from molecular sequences, Ann. Statist. 18 (1990), no. 2, 571–581.

[M] S. Mercier. Statistiques des scores pour l'analyse et la comparaison de séquences biologiques, Thèse de Doctorat de l'université de Rouen, 1999.

[N-W] S.B. Needleman & C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins, J. Mol. Biol. 48 (1970), 443–453.

[S-W] T.F. Smith & M.S. Waterman. Comparison of bio-sequences, Adv. Appl. Math. 2 (1981), 482–489.

[W-G-A] M.S. Waterman, L. Gordon & R. Arratia. Phase transitions in sequence matches and nucleic acid structure, Proc. Nat. Acad. Sci. U.S.A. 84 (1987), no. 5, 1239–1243.

[Z] Y. Zhang. A limit theorem for matching random sequences allowing deletions, Ann. Appl. Prob. 5 (1995), 1236–1240.