– Texte –
Recherche des zones codantes d'un brin ADN

Mots-clefs : chaîne de Markov, conditionnement, estimation d'une matrice de transition.

1 Structure (grossière) de l'ADN

Le génome d'un organisme vivant est constitué d'une ou quelques très longues molécules d'ADN. L'ADN est une molécule double constituée de deux brins. Chaque brin est lui-même constitué d'un enchaînement « désordonné » de nucléotides. Il y a quatre nucléotides : adénine, cytosine, guanine et thymine, notés respectivement a, c, g, t. Dans la double hélice, les nucléotides sont organisés en deux paires de lettres complémentaires a et t d'une part, c et g d'autre part. Cette redondance d'information permet la duplication sans altération du génome. Un brin d'ADN peut donc être vu comme un mot écrit avec l'alphabet $\mathcal{A} = \{a, c, g, t\}$.

Un brin d'ADN contient non seulement toute l'information nécessaire au développement des espèces vivantes (les gènes) mais aussi quantité de régions (dites intergéniques) qui semblent n'avoir aucune utilité fonctionnelle : nous appelerons respectivement ces zones codantes et non codantes. Les biologistes ont besoin d'algorithmes permettant de segmenter un brin en parties (vraisemblablement) codantes et non codantes. Le but de ce texte est de proposer un modèle de type markovien présentant différents régimes et un algorithme qui permet de reconnaître ces régimes.

2 Un modèle de Markov caché

Au brin d'ADN noté $X_1, \ldots, X_l$, on adjoint une suite $U_1, \ldots, U_l$ à valeurs dans $\mathcal{U} = \{0, 1\}$. La position i sera dite codante (resp. non codante) si $U_i = 1$ (resp. $U_i = 0$). La suite $(U_i)_i$ n'étant pas observée, la question est la suivante : peut-on se faire une idée de la valeur des $(U_i)_i$ à partir de celle des $(X_i)_i$?

Il faut pour cela poser un modèle plus précis. Dans la suite, nous supposerons que

- le processus complet $((X_n, U_n))_{n\geqslant 0}$ est une chaîne de Markov sur $\mathcal{A} \times \mathcal{U}$,
- le processus caché $(U_n)_{n\geqslant 0}$ est une chaîne de Markov sur $\mathcal{U}$ de matrice de transition ρ,
- il existe des matrices markoviennes $(\pi_u)_{u\in\mathcal{U}}$ telles que, pour tous $n \geqslant 0$, $u \in \mathcal{U}$, $x, x' \in \mathcal{A}$, $\mathbb{P}(X_{n+1} = x' | X_n = x, U_{n+1} = u) = \pi_u(x, x')$.

Nous supposerons de plus que toutes les matrices markoviennes ci-dessous sont à coefficients strictement positifs. Cette hypothèse est tout à fait naturelle dans la mesure où, dans la pratique, toute succession de deux nucléotides est possible.

Lemme 2.1. *La chaîne de Markov $(X_l, U_l)_{l\in\mathbb{N}}$ est homogène. Sa matrice de transition est donnée par*

$$P((x,u),(x',u')) := \mathbb{P}(X_{n+1}=x', U_{n+1}=u' | X_n = x, U_n = u) = \rho(u,u')\pi_{u'}(x,x').$$

Elle est de plus irréductible, récurrente et apériodique.

Remarque 2.2. Sauf exceptions (non intéressantes), le processus $(X_n)_n$ n'est pas une chaîne de Markov.

Remarque 2.3. La loi de la longueur d'un segment de la catégorie $u \in \mathcal{U}$ suit une loi géométrique de paramètre $1-\rho(u,u)$. Cette propriété markovienne est parfois limitante dans les applications.

Nous supposerons dans toute la suite que l'on connaît les paramètres du modèle. On pourra par exemple choisir les valeurs suivantes :

$$\rho = \begin{pmatrix} .99 & .01 \\ .02 & .98 \end{pmatrix}, \quad \pi_1 = \begin{pmatrix} .3 & .3 & .3 & .1 \\ .3 & .3 & .1 & .3 \\ .3 & .1 & .3 & .3 \\ .1 & .3 & .3 & .3 \end{pmatrix} \quad \text{et} \quad \pi_2 = \begin{pmatrix} .5 & .3 & .1 & .1 \\ .4 & .4 & .1 & .1 \\ .4 & .1 & .4 & .1 \\ .5 & .3 & .1 & .1 \end{pmatrix}$$

3 L'algorithme de segmentation

Notre but est ici de comprendre comment tirer le meilleur profit de l'observation d'une trajectoire de longueur l du processus X pour retrouver les divisions en catégories. Il s'agit donc de calculer, connaissant les matrices de transition, les probabilités

$$\forall v \in \mathcal{U}, \quad \mathbb{P}(U_i = v | X_1 = x_1, \ldots, X_l = x_l).$$

Remarque 3.1. Il existe des algorithmes permettant à la fois d'estimer les transitions ρ et π_u et de retrouver les divisions de la trajectoire du processus observé mais ils sont assez long à mettre en place.

On note :

- $P^i(v) := \mathbb{P}(U_i = v | X_1 = x_1, \ldots, X_{i-1} = x_{i-1})$ la probabilité de l'état caché v en position i sachant le passé de la chaîne observée jusqu'en $i-1$ (on parle de probabilité de prédiction),
- $F^i(v) := \mathbb{P}(U_i = v | X_1 = x_1, \ldots, X_i = x_i)$ la probabilité de l'état caché v en position i sachant le passé et le présent de la chaîne observée (on parle de probabilité de filtrage),
- $L^i(v) := \mathbb{P}(U_i = v | X_1 = x_1, \ldots, X_l = x_l)$ la probabilité de l'état caché v en position i sachant toute la trajectoire de la chaîne observée (on parle de probabilité de lissage).

L'algorithme que nous allons mettre en place est de type *forward-backward* : on calcule de proche en proche P^1, P^1,…,P^l et F^l, puis on en déduit (toujours de proche en proche mais en descendant) L^l,…,L^1. Les relations de récurrence nécessaires à cet algorithme sont rassemblées dans les deux propositions 3.2 et 3.3.

Proposition 3.2. *Les probabilités* $(P^i(v))_{1\leqslant i\leqslant l, v\in\mathcal{U}}$ *et* $(F^i(v))_{1\leqslant i\leqslant l, v\in\mathcal{U}}$ *vérifient les relations*

$$P^i(v) = \sum_{u\in\mathcal{U}} \rho(u,v) F^{i-1}(u), \tag{1}$$

$$F^i(v) = \frac{\pi_v(x_{i-1}, x_i) P^i(v)}{\sum_{u\in\mathcal{U}} \pi_u(x_{i-1}, x_i) P^i(u)}. \tag{2}$$

On initialise l'algorithme en choisissant pour $P^1(u)$ la loi initiale (par exemple la loi stationnaire). Les relations (1) et (2) permettent de calculer toutes les probabilités P^i et F^i.

Proposition 3.3. *Les probabilités de lissage vérifient les relations suivantes :*

$$L^{i-1}(u) = F^{i-1}(u) \sum_{v\in\mathcal{U}} \rho(u,v) \frac{L^i(v)}{P^i(v)}. \tag{3}$$

Démonstration. On montrera dans un premier temps que $\mathbb{P}(U_{i-1} = u, U_i = v | X_1 = x_1, \ldots, X_l = x_l)$ est égal à

$$\frac{\rho(u,v)\mathbb{P}(U_{i-1} = u | X_1 = x_1, \ldots, X_{i-1} = x_{i-1})\mathbb{P}(U_i = v | X_1 = x_1, \ldots, X_l = x_l)}{\mathbb{P}(U_i = v | X_1 = x_1, \ldots, X_{i-1} = x_{i-1})},$$

puis on en déduit la relation (3). □

4 Markov ou pas ?

L'étude précédente suppose connues les matrices de transition. Une question naturelle est de savoir comment estimer la matrice de transition d'une chaîne de Markov à partir d'une de ses trajectoires. Une question encore plus cruciale serait de pouvoir décider si la trajectoire que l'on observe peut être onsidérer comme une chaîne de Markov. L'objet de cette section est d'apporter des réponses à ces deux questions.

4.1 Estimation d'une matrice de transition

On considère une chaîne de Markov canonique $\left(\Omega, \mathcal{A}, (\mathbb{P}_x)_{x\in E}, (X_n)_{n\in\mathbb{N}}\right)$ associée à la matrice de transition $\Pi = (\Pi(i,j))_{(i,j)\in E^2}$. On suppose que E est fini de cardinal s et que la chaîne est irréductible et récurrente. Notons μ la mesure invariante de la chaîne. On note pour $(i,j) \in E^2$,

$$N_n^i = \sum_{p=0}^{n-1} \mathbf{1}_{\{X_p = i\}} \quad \text{et} \quad N_n^{ij} = \sum_{p=0}^{n-1} \mathbf{1}_{\{X_p = i, X_{p+1} = j\}}.$$

On note enfin $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ et $\mathbb{F} = (\mathcal{F}_n)_n$. Supposons la mesure initiale ν connue mais la matrice de transition Π inconnue.

Notons λ la mesure de comptage sur E. La loi du $(n+1)$-uplet $(X_0, X_1, \ldots, X_n)$ admet pour densité par rapport à la mesure $\lambda^{\otimes n+1}$ la fonction f_n définie par

$$\begin{aligned} f_n(\Pi, x_0, x_1, \ldots, x_n) &= \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) \\ &= \nu(x_0)\Pi(x_0, x_1) \cdots \Pi(x_{n-1}, x_n). \end{aligned}$$

La vraisemblance $L_n(\Pi)$ de l'échantillon $(X_0, X_1, \ldots, X_n)$ est la valeur de f_n prise au point $(X_0, X_1, \ldots, X_n)$. La log-vraisemblance est définie par $l_n(\Pi) = \ln L_n(\Pi)$. Trouver un estimateur du maximum de vraisemblance revient à trouver une matrice markovienne qui maximise $l_n(\Pi)$.

Proposition 4.1. *L'estimateur de maximum de vraisemblance de* Π *est donné par*

$$\textit{pour } 1 \leqslant i, j \leqslant s, \quad \widehat{\Pi}_n(i,j) = \begin{cases} \dfrac{N_n^{ij}}{N_n^i} & \textit{si } N_n^i > 0, \\ 0 & \textit{sinon.} \end{cases}$$

Démonstration. Comme la somme de chaque ligne de la matrice Π doit être égale à 1, le véritable paramètre est $\{(\Pi(i,1), \ldots, \Pi(i,s-1)),\ 1 \leqslant i \leqslant s\}$ élément de $\mathbb{R}^{s(s-1)}$. On peut réécrire la log-vraisemblance comme :

$$l_n(\Pi) = \sum_{i=1}^{s} \left[\sum_{j=1}^{s-1} N_n^{ij} \ln \Pi(i,j) + N_n^{is} \ln \left(1 - \sum_{j=1}^{s-1} \Pi(i,j) \right) \right].$$

Reste à déterminer les extrema de cette fonction... □

Quelles sont les propriétés asymptotiques de cet estimateur ? Elles se déduisent du comportement des suites $(N_n^{ij})_n$ et $(N_n^i)_n$ qui sont établies ci-dessous.

Théorème 4.2. *Pour tout* $x \in E$ *et tout* $(i,j) \in E^2$,

$$\frac{1}{n} N_n^i \xrightarrow[n\to\infty]{\mathbb{P}_x - p.s.} \mu(i), \quad \frac{1}{n} N_n^{ij} \xrightarrow[n\to\infty]{\mathbb{P}_x - p.s.} \mu(i)\Pi(i,j)\ ;$$

$$\frac{1}{\sqrt{n}} \big(N_n^{ij} - N_n^i \Pi(i,j) \big) \xrightarrow[n\to\infty]{\mathcal{L}(\mathbb{P}_x)} \mathcal{N}(0, \mu(i)\Pi(i,j)(1 - \Pi(i,j))).$$

Démonstration. La première convergence presque sûre découle du théorème ergodique pour X. Pour la seconde, il faut faire les remarques suivantes. On peut construire à partir de X la suite Y à valeurs dans E^2 définie par $Y_n = (X_n, X_{n+1})$. On constate alors que Y est elle-même une chaîne de Markov irréductible récurrente de mesure invariante $\mu(i)\Pi(i,j)$.

Les résultats de convergence en loi seront admis. □

On déduit de ce théorème les propriétés de convergence de l'estimateur $\widehat{\Pi}$.

Corollaire 4.3. *Pour tout $x \in E$ et tous $i, j \in E$,*

$$\widehat{\Pi}_n(i,j) \xrightarrow[n\to\infty]{\mathbb{P}_x - p.s.} \Pi(i,j) \ ;$$

$$\sqrt{n\mu(i)}(\widehat{\Pi}_n(i,j) - \Pi(i,j)) \xrightarrow[n\to\infty]{\mathcal{L}(\mathbb{P}_x)} \mathcal{N}(0, \Pi(i,j)(1-\Pi(i,j))).$$

Remarque 4.4. D'après l'hypothèse de récurrence, $\Pi(i,j)$ est strictement inférieur à 1 mais peut être nul, auquel cas l'estimateur de $\Pi(i,j)$ est excellent...

4.2 Un test de markovianité

On voudrait à présent pouvoir décider si une suite finie de variables aléatoires peut raisonnablement être considérée comme une chaîne de Markov. Le théorème suivant nous fournit l'outil-clé pour cela.

Théorème 4.5. *Soit $(X_l)_{l\geqslant 1}$ une chaîne de Markov irréductible sur E fini de cardinal s et de matrice de transition strictement positive. On note, pour i, j et k dans E,*

$$N_l^i = \sum_{n=1}^{l} \mathbf{1}_{\{X_n=i\}}, \quad N_l^{ij} = \sum_{n=1}^{l-1} \mathbf{1}_{\{X_n=i, X_{n+1}=j\}} \quad \text{et} \quad N_l^{ijk} = \sum_{n=1}^{l-2} \mathbf{1}_{\{X_n=i, X_{n+1}=j, X_{n+2}=k\}}.$$

Alors

$$Z_l = \sum_{(i,j,k)\in E^3} \frac{(N_l^{ijk} - N_l^{ij}N_l^{jk}/N_l^j)^2}{N_l^{ij}N_l^{jk}/N_l^j} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(s(s-1)^2).$$

Ce théorème qui sera admis et que l'on ne cherchera pas à démontrer permet de mettre en place un test de type χ^2 pour décider de la markovianité d'une suite.

5 Suggestions

1. On pourra détailler la modélisation d'un brin d'ADN par un modèle de Markov caché.
2. On pourra démontrer le lemme 2.1 et exprimer la mesure invariante de $(X_n, U_n)_n$ en fonction des paramètres.
3. On pourra expliciter, théoriquement et par la simulation, le comportement asymptotique (quand l tend vers l'infini) de $(1/l)\sum_{i=1}^{l} \mathbf{1}_{\{X_i=k\}}$ pour $k \in \mathcal{A}$ en fonction des paramètres du modèle.
4. On pourra démontrer la proposition 3.2.
5. On pourra démontrer la proposition 3.3.

6. On pourra illustrer par la simulation l'efficacité de l'algorithme en estimant notamment la proportion de sites correctement annotés dans l'exemple donné dans le texte. On pourra également faire varier les paramètres, et en premier lieu, la matrice de transition ρ.
7. On pourra expliquer comment obtenir l'estimateur du maximum de vraisemblance de la matrice de transition Π.
8. On pourra démontrer les convergences presque sûres du théorème 4.2.
9. On pourra illustrer par la simulation le corollaire 4.3.
10. On pourra utiliser le théorème 4.5 pour illustrer la remarque 2.2 ; à savoir que la partie observée X du processus complet (X, U) n'est (en général) pas une chaîne de Markov.

