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

## 1 Position du problème

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 $\mathcal{A} = \{a, c, g, t\}$.

Le modèle le plus simpliste pour modéliser la répartition des lettres dans une séquence est appelé *modèle de Bernoulli*. On modélise une séquence de longueur $l$ par une suite $(X_i)_{1 \leqslant i \leqslant l}$ de variables aléatoires indépendantes identiquement distribuées de loi $\mu$ (inconnue) sur $\mathcal{A}$. Ce cadre est généralement très insuffisant pour rendre compte des relations entre lettres successives. On raffine donc classiquement en supposant que la suite $(X_i)_{1 \leqslant i \leqslant l}$ forme une chaîne de Markov homogène sur $\mathcal{A}$ de matrice de transition $\Pi$ (inconnue) que l'on supposera à coefficients strictement positifs. Cette hypothèse ne semble pas du tout restrictive puisqu'en pratique n'importe quelle lettre peut succéder à une autre.

**Exemple 1.1.** Sur une séquence de 1000 nucléotides que l'on a regroupés en deux classes, 1 pour les purines ($c$ et $g$) et 2 pour les pyrimides ($a$ et $t$), on a relevé les résultats suivants :

$$N^1 = 527, \quad N^2 = 473,$$

$$N^{11} = 241, \quad N^{12} = 286, \quad N^{21} = 285, \quad N^{22} = 187,$$

$$N^{111} = 115, \quad N^{112} = 126, \quad N^{121} = 172, \quad N^{122} = 113$$

$$N^{211} = 126, \quad N^{212} = 159, \quad N^{221} = 113 \quad \text{et} \quad N^{222} = 74,$$

où $N^{ijk}$ est le nombre d'occurences de la suite $ijk$ dans le brin.

Puisque $\pi_{ij} = \mathbb{P}(X_1 = j | X_0 = i)$, $\pi_{ij}$ peut être estimé par $\hat{\pi}_{ij} = N^{ij}/N^i$. D'après les données de l'exemple 1.1, on en déduit donc l'estimation suivante pour la matrice de transition :

$$\hat{\Pi} = \begin{pmatrix} 0.457 & 0.543 \\ 0.603 & 0.397 \end{pmatrix} \tag{1}$$

Le modèle, une fois les paramètres estimés, prévoit donc par exemple que

$$\mathbb{P}(X_2 = 1, X_1 = 1 | X_0 = 1) = 0.209 \quad \text{et} \quad \mathbb{P}(X_2 = 2, X_1 = 1 | X_0 = 1) = 0.248,$$

alors que les estimations de ces probabilités à partir des relevés sur le brin donnent respectivement 0.218 et 0.239. Pour décider si ces écarts sont susceptibles de s'expliquer par l'aléa du modèle ou s'ils sont le signe d'une inadéquation du modèle, on peut s'appuyer sur le théorème suivant.

**Theorème 1.2.** *Soit $(X_l)_{l \geqslant 1}$ une chaîne de Markov récurrente sur $E$ fini de cardinal $s$ et de matrice de transition strictement positive. On note, pour $i$, $j$ et $k$ dans $E$,*

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

*Alors*

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

*Remarque* 1.3. En confrontant les mesures et le théorème 1.2, le modèle markovien semble bien inadapté. Cela signifie (vraisemblablement) qu'un brin d'ADN est agencé par un mécanisme plus complexe que celui d'une chaîne de Markov simple.

Dans un brin d'ADN, certains parties, les gènes, sont codantes, c'est-à-dire qu'elles sont utilisées par la cellule pour synthétiser les protéines dont elle a besoin ; d'autres, les zones intergéniques, sont non codantes et leur utilité n'est pas toujours bien comprise par les scientifiques[1]. Cette hétérogénéité rend le modèle de chaîne de Markov homogène souvent inapproprié pour modéliser de longs brin d'ADN. Les biologistes ont donc 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

Dans chacune des parties (codante et non codante) de l'ADN, on propose de modéliser la succession des lettres par une chaîne de Markov mais, pour rendre compte de l'hétérogénéité de l'ADN, les transitions dépendront de la région dans laquelle se trouve la chaîne. On parle alors de processus de Markov caché. Il se compose de deux processus, l'un caché, l'autre observé.

Le premier, le processus caché $(U_n)_{n\in\mathbb{N}^*}$, est une chaîne de Markov sur l'espace $\mathcal{U}$ des catégories possibles. Ce processus représente la segmentation sous-jacente de la séquence : une suite ininterrompue du même état caché $u$ indique un segment de cette catégorie (par exemple *codant* ou *non codant*).

Le second, le processus observé $(X_n)_{n\in\mathbb{N}^*}$ à valeurs dans $\mathcal{A}$, représente la séquence tel que le couple $(X_n, U_n)_n$ soit une chaîne de Markov et, conditionnellement au processus caché, le processus observé soit une chaîne de Markov hétérogène (*i.e.* dont la matrice de transition dépend de du temps $n$).

Plus précisément, la loi du processus $(X_n, U_n)$ est définie par les relations suivantes :

$$\begin{aligned}
&\mathbb{P}(X_1 = x_1, \dots, X_l = x_l, U_1 = u_1, \dots, U_l = u_l) \\
&\quad = \mathbb{P}(U_1 = u_1, \dots, U_l = u_l)\mathbb{P}(X_1 = x_1, \dots, X_l = x_l | U_1 = u_1, \dots, U_l = u_l) \\
&\quad = \nu(u_1) \prod_{i=2}^{l} \rho(u_{i-1}, u_i)\mu_{u_1}(x_1) \prod_{i=2}^{l} \pi_{u_i}(x_{i-1}, x_i),
\end{aligned}$$

où $(\rho(u,v))_{u,v\in\mathcal{U}}$ est la matrice de transition de $U$, $\nu$ sa mesure initiale, pour tout $u \in \mathcal{U}$, $(\pi_u(i,j))_{i,j\in\mathcal{A}}$ est une matrice de transition sur $\mathcal{A}$ et $\mu_u$ est la mesure initiale de $X$ sachant que $U_1 = u$. Ce sont les paramètres du modèle. On supposera dans la suite que toutes les matrices de transitions $\rho$ et $(\pi_u)_{u\in\mathcal{U}}$ sont à coefficients strictement positifs.

**Lemme 2.1.** *Le processus complet $(X_l, U_l)_{l\in\mathbb{N}}$ est une chaîne de Markov homogène sur $\mathcal{A}\times\mathcal{U}$. Elle est 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.

On prend les valeurs suivantes pour les paramètres du modèle :

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

Nous supposerons dans toute la suite que l'on connaît les paramètres du modèle.

## 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 (par exemple codantes et non codantes). Il s'agit donc de calculer, connaissant les matrices de transition, les probabilités

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

[1] La structure réelle d'un brin d'ADN est en fait bien plus complexe. Les gènes peuvent être découpés en exons, les segments qui seront traduits, et en introns, les segments qui seront éliminés lors de la maturation de l'ARN messager, avant traduction...

 

*Remarque* 3.1. Il existe des algorithmes permettant à la fois d'estimer les transitions $\rho$ et $\pi_u$ 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*

$$P^i(v) = \sum_{u \in \mathcal{U}} \rho(u,v) F^{i-1}(u), \tag{2}$$

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

On initialise l'algorithme en choisissant pour $P^1(u)$ la loi initiale (par exemple la loi stationnaire). Les relations (2) et (3) permettent de calculer toutes les probabilités $P^i$ et $F^i$.

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

$$L^{i-1}(u) = F^{i-1}(u) \sum_{v \in \mathcal{U}} \rho(u,v) \frac{L^i(v)}{P^i(v)}. \tag{4}$$

*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 (4). □

## 4 Suggestions

1. On pourra présenter les avantages et inconvénients du modèle de chaîne de Markov homogène. En particulier, on pourra s'appuyer pour cela sur l'exemple 1.1 et justifier l'estimation (1) de la matrice de transition.
2. On pourra expliciter le raisonnement suggéré par la remarque 1.3 en décrivant une méthode statistique permettant, grâce au théorème 1.2, de justifier la phrase *au vu de ces données, le modèle markovien semble bien improbable.*
3. On pourra détailler la modélisation d'un brin d'ADN par un modèle de Markov caché.
4. On pourra démontrer le lemme 2.1 et exprimer la mesure invariante de $(X_n, U_n)_n$ en fonction des paramètres.
5. 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.
6. On pourra commenter et approfondir les remarques 2.2 et 2.3.
7. On pourra démontrer la proposition 3.2.
8. On pourra démontrer la proposition 3.3.
9. 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 $\rho$.

 