Texte (oraux blancs)
Détection de segments inversés dans une séquence ADN

Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury demande que la discussion soit accompagnée d'exemples traités sur ordinateur. Il est souhaitable que vous organisiez votre présentation comme si le jury n'avait pas connaissance du texte. Le jury aura néanmoins le texte sous les yeux pendant votre exposé.

1 Retournement dans les sequences ADN

La molécule d'ADN est composée de deux brins, chacun étant une séquence linéaire et orientée de bases, ou nucléotides : il y a quatre bases, représentées par une lettre : l'adénine (a), la cytosine (c), la guanine (g) et la thymine (t). L'ADN peut se représenter comme une suite de lettres dans l'alphabet $\mathcal{A} = \{a, c, g, t\}$. Ces bases sont deux à deux complémentaires : (a) et (t) d'une part et (c) et (g) d'autre part se lient l'une à l'autre par des liaisons hydrogène. La molécule est donc constituée de deux brins complémentaires, formés par un long squelette sur lequel s'attachent une suite de nucléotides. Chacun de ces nucléotides se lie au nucléotide correspondant, sur l'autre brin, par liaison hydrogène.

Les séquences biologiques sont altérées au cours du temps. Au dela des transformations classiques de la séquence par substitution (nucléotide remplacé par un autre), par insertion de nouveaux nucléotides et délétion de nucléotides, il existe des remaniements globaux qui peuvent jouer des rôles importants dans l'évolution d'une séquence. Parmi eux, l'inclusion après retournement d'un morceau de séquence provenant du brin complémentaire que l'on appelera par suite segment inversé ou *Dincom* (DNA inverse complementary). L'objet de ce texte est de proposer une approche permettant de détecter la présense de Dincom dans une séquence ADN.

2 Modélisation markovienne

On modélise assez couramment une séquence ADN de longueur n, $x_1^n = (x_1, \cdots, x_n) \in \mathcal{A}^n$ comme la réalisation d'un processus $X_1^n = (X_i)_{1 \leqslant i \leqslant n}$ définie une chaîne de markov stationnaire de longueur finie de matrice de transition Q_0 *strictement positive* et de probabilité invariante μ_0 ie

$$P(X_i = x_i,\ 1 \leqslant i \leqslant n) = \mu_0(x_1) \prod_{i=1}^{n-1} Q_0(x_i, x_{i+1}).$$

Si $x \to \overline{x}$ est l'opération de complémentation sur $\mathcal{A}$ (définie par $\overline{a} = t$, $\overline{t} = a$, $\overline{c} = g$ et $\overline{g} = c$), alors l'opération de retournement et de complémentation définit un nouveau processus $Y_1^n = (\overline{X}_{n+1-i})_{1 \leqslant i \leqslant n}$.

Proposition 1. Le processus $Y_1^n = (\overline{X}_{n+1-i})_{1\leqslant i\leqslant n}$ est également une chaîne de Markov de longueur finie de matrice de transition Q_1 et de loi invariante μ_1 vérifiant

$$\mu_1(x) = \mu_0(\overline{x}),\ Q_1(x,x') = Q_0(\overline{x'},\overline{x})\frac{\mu_0(\overline{x'})}{\mu_0(\overline{x})} \tag{1}$$

pour tout $x, x' \in \mathcal{A}$.

3 Construction d'un test de détection

En notant

$$p_{\epsilon,l}(x_1,\cdots,x_l) \doteq \mu_\epsilon(x_1)\prod_{i=1}^{l-1} Q_\epsilon(x_i,x_{i+1})\,, \tag{2}$$

on veut tester si un fragment $x_p^q = (x_p,\cdots,x_q)$ de longueur l, $(1\leqslant p\leqslant q\leqslant n,\ q = p+l-1)$ est un Dincom à partir d'un test d'hypothèses simples entre H_0 : "x_p^q est généré selon la distribution $p_{0,l}$" contre H_1 : "x_p^q est généré selon la distribution $p_{1,l}$". Pour cela, on introduit la statistique du rapport de vraisemblance :

$$T_l(x_p^q) = \log(\frac{p_{1,l}(x_p^q)}{p_{0,l}(x_p^q)}) \tag{3}$$

pour laquelle on a

$$Q_0 \neq Q_1 \implies E_0(T_l) < E_1(T_l)\,, \tag{4}$$

où $E_\epsilon(T_l) = \sum_{x_1^l\in\mathcal{A}^l} T_l(x_1^l)p_{\epsilon,l}(x_1,\cdots,x_l)$ pour tout $\epsilon \in \{0,1\}$.

Un test optimal de H_0 contre H_1 s'obtient en rejetant H_0 lorsque $T_l(x_p^q)$ est au-dessus d'un seuil.

Cependant, on s'intéresse plutôt au test simultané de tous les fragments de longueur l commençant à une position p quelconque c'est-à-dire que H_0 est maintenant "x_1^n est généré selon la distribution $p_{0,n}$" et H_1 devient : "il existe $1\leqslant p\leqslant n-l+1$ tel que x_p^{p+l-1} est généré selon la distribution $p_{1,l}$"

Il s'agit donc d'un test multiple que l'on construit sur la statistique

$$S_l(x_1^n) \doteq \max_{1\leqslant p\leqslant n-l+1} T_l(x_p^{p+l-1})\,. \tag{5}$$

Le rejet de H_0 est obtenu maintenant lorsque S_l dépasse un certain seuil s_l à définir : $d_l = \mathbf{1}_{S_l>s_l}$.

4 Calibration directe de d_l

Le problème est alors de choisir s_l pour obtenir un test de niveau $\alpha \in]0,1[$ prescrit. Dans les situations concrètes, on aura toujours $l << n$. Comme la loi de S_l n'est généralement pas explicite, le calcul de s_l demande un traitement spécial.

Une première approche consiste à estimer la fonction de répartition $F_l(s) \doteq P(S_l(X_1^n)\leqslant s)$ ou encore de survie $G_l(s) \doteq 1-F_l(s)$ par une méthode de Monte Carlo. On simule alors un N-échantillon $(Z_1,\cdots,Z_N)$ de v.a.i.i.d de loi commune la loi de $S_l(X_1^n)$ puis on estime s_l par

$$\hat{s}_l = Z_{\sigma(k_*)} \tag{6}$$

où $1 \leqslant k_* \leqslant N$ est le plus petit entier tel que $k/N \geqslant 1-\alpha$ et $Z_{\sigma(1)} \leqslant \cdots \leqslant Z_{\sigma(N)}$ est le ré-arrangement des Z_i par ordre croissant. Cette procédure est basée sur l'estimation de la fonction de répartition des Z_i par la fonction de répartition empirique.

Une autre façon de calibrer le test est de remarquer que sous P, les variables $T_l(X_p^{p+l-1})$ ont même loi pour toutes les valeurs de p, si bien que

$$E(d_l(x_1^n)) \leqslant nP(T_l(X_1^l) > s_l)\,. \tag{7}$$

Il suffit de chercher s_l tel que $P(T_l(X_1^l) > s_l) \leqslant \alpha/n$ pour que d_l soit de niveau α. On peut à nouveau procéder par Monte Carlo. Cependant comme les valeurs de s intéressantes sont celles pour la probabilité de l'événement à estimer est très faible, beaucoup d'itérations de Monte Carlo seront nécessaires pour estimer correctement la queue de la distribution.

5 Echantillonage pondéré

Pour accélerer l'estimation, on peut utiliser un échantillonnage pondéré. On simule alors dans cette deuxième approche un N-échantillon pondéré $((Z'_1, W_1), \cdots, (Z'_N, W_N))$ de v.a.i.i.d. de loi commune la loi de $(T_l(Y_1^l), \frac{p_{0,l}}{p_{1,l}}(Y_1^l))$ lorsque Y_1^l suit la loi p_1^l, et on estime s_l par

$$\hat{s}_l = Z'_{\sigma(k_*)} \tag{8}$$

où $1 \leqslant k_* \leqslant N$ est le plus grand entier k tel que $\sum_{i=1}^{k-1} W_{\sigma(i)}/N \leqslant \alpha/n$ et $Z'_{\sigma(1)} \leqslant \cdots \leqslant Z'_{\sigma(N)}$ est le ré-arrangement des Z'_i par ordre *décroissant*.

6 Suggestions pour le développement

-Commenter la modélisation faite pour la sequence d'ADN et les modélisations alternatives éventuelles.

-Montrer la proposition 1

-Dans la partie 3 du texte, montrer l'implication (4) et expliquer en quel sens le seuillage sur $T_l(x_p^q)$ est optimal pour le test d'hypothèses simples.

-Expliquer le principe de la construction du test multiple.

-Expliquer le comportement de $\hat{s}_l$ dans le cas de la première procédure de calibration lorsque N tend vers l'infini.

-Mettre en oeuvre la première approche de calibration de d_l pour $n = 1000$ et $l = 100$ dans le dans le cas du génome du HIV1 pour lequel la matrice de transition Q_0 est

$$Q_0 = \begin{pmatrix} 0.326 & 0.300 & 0.165 & 0.209 \\ 0.346 & 0.279 & 0.193 & 0.182 \\ 0.448 & 0.054 & 0.233 & 0.265 \\ 0.316 & 0.273 & 0.158 & 0.253 \end{pmatrix} \tag{9}$$

où les colonnes représentent les transitions vers a, c, g et t (dans cet ordre de gauche à droite) et de même pour les lignes de haut en bas. On trouve pour la distribution invariante $\mu_0 = (0.351, 0.244, 0.182, 0.223)$.
-Expliquer pourquoi une méthode de Monte Carlo classique ne peut convenir pour la deuxième approche.
-Expliquer le principe de l'échantillonnage pondéré dans la deuxième approche et la mettre en oeuvre pour le HIV1. Le seuil calculé par la deuxième approche est notablement plus grand que par la première. Expliquer pourquoi ainsi que l'intérêt éventuel de choisir l'une ou l'autre méthode (rapidité des calculs, robustesse des tests, comportement vis à vis d'hypothèses alternatives).

