Variables aléatoires à densité

Problème :

Notations : Dans tout le problème, $\mathbb{R}$ désigne l'ensemble des nombres réels et n **désigne un entier naturel supérieur ou égal à** 2 **fixé**.

Enfin $\mathbb{I}_{]a,b[}(x)$ désignera la fonction indicatrice sur $]a,b[$ définie par $\mathbb{I}_{]a,b[}(x) = \begin{cases} 1 & \text{si } a < x < b \\ 0 & \text{sinon} \end{cases}$.

Rappel : Si $(X_1, \cdots, X_n)$ sont n variables aléatoires indépendantes qui suivent toutes une même loi de bernoulli de paramètre p, alors on dira que $S_n = X_1 + \cdots + X_n$ est la variable aléatoire égale au nombre de succès au cours de n épreuves de bernoulli indépendantes et S_n suit une loi binomiale de paramètres n et p.

Réarrangements chromosomiques

Introduction :

On appelle **réarrangements évolutifs** les réarrangements qui différencient les génomes d'espèces différentes. Il se sont produits dans des cellules de la lignée germinale et ont été transmis à la descendance, se fixant dans la population, soit par dérive, soit par sélection naturelle. Ils ont été mis en évidence dès 1921 par Sturtevant grâce à une inversion sur un chromosome de drosophile et depuis, les génomes d'espèces différentes ont été comparés. Avec les travaux de Nadeau et Taylor sur la souris, en 1984, il apparaît que, non seulement l'essentiel de gènes humains ont leur contrepartie chez la souris, mais que des gènes similaires se succèdent dans un même ordre chez les deux espèces et sur des segments de gènes plus ou moins longs (on parlera de « blocs synthéniques »). A cette date, en regardant à grand échelle les génomes respectifs de l'homme et de la souris, il a été envisagé comme possible de passer de l'un à l'autre par seulement 7 inversions (cf. annexe).

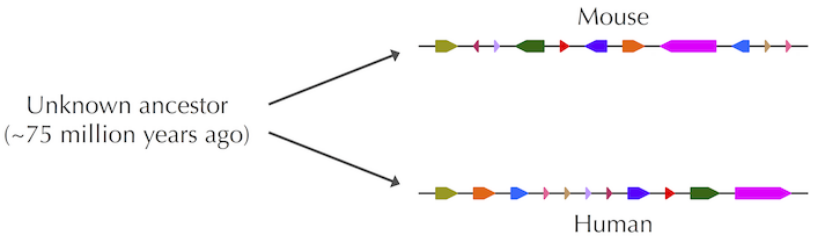

FIGURE 1 – `Chromosomes X de souris et d'humains en 7 segments orientés et colorés (blocs synthéniques)`

Les mécanismes moléculaires responsables de la formation des réarrangements chromosomiques sont bien compris aujourd'hui. En revanche il reste de nombreuses inconnues quant à ce qui détermine leur fréquence et leurs localisation dans le génome. En particulier, l'une des questions en suspend des études sur l'évolution chromosomique est de savoir s'il existe sur le génome des régions plus fragiles, des « lignes de faille », qui expliqueraient ces cassures ou bien si le processus est entièrement aléatoire, les fractures étant indépendantes et uniformément distribuées le long du génome.

L'objectif de cette première partie est d'utiliser les variables aléatoires à densité pour discuter cette dernière hypothèse qui a donné lieu au modèle dit des « de cassures aléatoires » par opposition au modèle « des régions fragiles ».

Présentation du modèle.

D'un point de vue algorithmique, on visualisera une inversion comme une double cassure sur le génome, créant ainsi un intervalle au sein duquel le segment est inversé avant d'être recollé. [1]. Pour créer p inversions, on placera ainsi aléatoirement de façon uniforme et indépendante $n = 2 * p$ cassures qui engendreront $n + 1$ segments. Un tel réarrangement sera donc modélisé par une liste de n cassures numérotées de 1 à n.
La taille du génome humain étant de l'ordre de $N = 3.4e9$ bases (soit 3.4 Gb), **on peut modéliser le lieu d'une cassure** i $(1 \leq i \leq n)$ soit par une variable aléatoire discrète $C_i \hookrightarrow \mathcal{U}_{[\![1,3.4e9]\!]}$ ou bien, au regard du nombre presque infini des valeurs pouvant être prises par C_i, **par une variable aléatoire à densité** $X_i = \dfrac{C_i}{N}$ **dont on considérera qu'elle suit une loi uniforme sur le segment** $]0,1[$. **Les variables** X_i **sont par hypothèse mutuellement indépendantes.**

Exemple : Obtenir $X_1 = 0.351$ et $X_2 = 0.10322563$ signifie que la première et la seconde cassures ont lieu respectivement à la place $0.351 * 3.4e9 = 1193400000$ et à la place $0.10322563 * 3.4e9 = 350967142$.

On note enfin $(Y_1, Y_2, \cdots, Y_n)$ les n variables aléatoires ayant pour valeur les mêmes valeurs que les variables $(X_1, X_2, \cdots, X_n)$ ordonnées dans l'ordre croissant. Ce sont elles qui donnent accès aux longueurs des segments inversés. Par exemple, pour $n = 4$, si on obtient $X_1 = 0.3$, $X_2 = 0.1$, $X_3 = 0.7$ et $X_4 = 0.2$, on aura $Y_1 = 0.1$, $Y_2 = 0.2$, $Y_3 = 0.3$ et $Y_4 = 0.7$ et des segments de longueurs $L_1 = Y_2 - Y_1 = 0.1$, $L_2 = Y_3 - Y_2 = 0.1$ et $L_3 = Y_4 - Y_3 = 0.4$.

Première partie : Approche mathématique.

1. On s'intéresse tout d'abord aux sites de la première et de la dernière cassure, à savoir respectivement $Y_1 = min\{X_1, ..., X_n\}$ et $Y_n = max\{X_1, ..., X_n\}$.
   a) Donner une densité f des variables aléatoires X_i ainsi que leur fonction de répartition F_X. Calculer leur espérance et leur variance.
   b) Écrire une fonction Python `simulCassure(p)` permettant de retourner une liste `Lc` de $n = 2 * p$ cassures prises au hasard dans $]0,1[$ selon le modèle des cassures aléatoires associé à des lois uniformes sur l'intervalle $]0,1[$.
   c) Montrer que pour tout x dans $]0,1[$, $\mathbb{P}(Y_n \leq x) = x^n$.
   d) Donner la fonction de répartition F_n de Y_n. En déduire que Y_n est une variable aléatoire à densité. Montrer que la fonction f_n définie sur $\mathbb{R}$ par $f_n(x) = nx^{n-1}\mathbb{I}_{]0,1[}(x)$ est une densité de Y_n.
   e) Montrer que pour tout x dans $]0,1[$, $\mathbb{P}(Y_1 > x) = (1-x)^n$.
   f) Donner la fonction de répartition F_1 de Y_1. En déduire que Y_1 est une variable aléatoire à densité. Montrer que la fonction f_1 définie sur $\mathbb{R}$ par $f_1(x) = n(1-x)^{n-1}\mathbb{I}_{]0,1[}(x)$ est une densité de Y_1.

2. a) Montrer que Y_n et $1 - Y_1$ ont la même loi. En déduire des relations entre $\mathbb{E}(Y_1)$ et $\mathbb{E}(Y_n)$ d'une part, $\mathbb{V}(Y_1)$ et $\mathbb{V}(Y_n)$ d'autre part.
   b) Montrer que $\mathbb{E}(Y_n) = \dfrac{n}{n+1}$. En déduire $\mathbb{E}(Y_1)$.
   c) Montrer que $\mathbb{V}(Y_n) = \dfrac{n}{(n+1)^2(n+2)}$. En déduire $\mathbb{V}(Y_1)$.
   d) Écrire une fonction Python `loiY1(m,p)` qui réalise m fois (avec m grand) la création d'une liste de $2 * p$ cassures et retourne la liste `LY1` des valeurs prises par Y_1 ainsi que sa moyenne et sa variance dont on rappellera la définition, estimations statistiques de $\mathbb{E}(Y_1)$ et $\mathbb{V}(Y_1)$.

1. cf. document en annexe fourni à titre indicatif

3. Suite à $n = 2 * p$ cassures aléatoires, on s'intéresse à la distribution des longueurs des segments obtenus. On va notamment s'arrêter sur les k premières cassures où k est un entier inférieur à n suffisamment grand et on cherchera la loi et l'espérance de Y_k, $1 \leq k \leq n$, site d'arrivée de la k-ième cassure.
   Pour tout réel $x \in [0,1]$ fixé, on introduit pour tout $i \in [\![1,n]\!]$ la variable aléatoire U_i valant 1 si l'événement $(X_i \leq x)$ est réalisé, 0 sinon. On introduit également la variable aléatoire $U = \sum_{i=1}^{n} U_i$.

   a) Justifier le fait que $U \hookrightarrow \mathcal{B}(n,x)$, loi binomiale de paramètres n et x.

   b) Pour un entier naturel k $(1 \leq k \leq n)$ fixé, justifier que les événements $(Y_k \leq x)$ et $(U \geq k)$ sont égaux. En déduire que $\mathbb{P}(Y_k \leq x) = \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j}$.

4. Dans la suite de cette partie, on notera F_k la fonction de répartition de Y_k et f_k sa densité continue sur $\mathbb{R}$ pour tout $k \in [\![2, n-1]\!]$ (*on ne cherchera pas à calculer* f_k).

   a) Pour tout $x \in [0,1]$ et $k \in [\![2, n-1]\!]$, montrer que $F_k(x) = 1 - F_{n-k+1}(1-x)$. En déduire que $f_k(x) = f_{n-k+1}(1-x)$ puis une relation entre $\mathbb{E}(Y_k)$ et $\mathbb{E}(Y_{n-k+1})$ en pensant au changement de variable $y = 1-x$.
   Montrer de même que $\mathbb{E}\left[(1-Y_k)^2\right] = \mathbb{E}(Y_{n-k+1}^2)$ et en déduire que $\mathbb{V}(Y_k) = \mathbb{V}(Y_{n-k+1})$.

   b) Montrer que pour tout $k \in [\![1,n]\!]$, $\mathbb{E}(Y_k) = \int_0^1 [1 - F_k(x)]dx$. En déduire que l'espérance de la longueur du k-ième segment $(1 \leq k \leq n-1)$ vaut : $\mathbb{E}(Y_{k+1} - Y_k) = \binom{n}{k} \int_0^1 x^k (1-x)^{n-k} dx$.

   c) On introduit la suite de terme général $I_k(n) = \binom{n}{k} \int_0^1 x^k (1-x)^{n-k} dx$ pour tout $k \in [\![0,n]\!]$.
   Montrer que $I_k(n) = I_{k+1}(n)$, $\forall k \in [\![0, n-1]\!]$.
   En déduire que $I_k(n) = \mathbb{E}(Y_1)$ pour tout entier naturel k, $0 \leq k \leq n$.

   d) Montrer que pour tout entier k $(1 \leq k \leq n)$, $\mathbb{E}(Y_k) = \dfrac{k}{n+1}$. Interpréter.

Seconde partie : Approche algorithmique.

Les travaux sur l'évolution du génome ont montré que les longueurs des segments obtenus par cassures présentent une grande diversité (de quelques Kb à plusieurs dizaines de Mb). Le modèle des cassures aléatoires rend-il compte de cette forte variance ?

1. Montrer que, si n est suffisamment grand, il est possible d'approcher la loi de Y_1 par une loi exponentielle de paramètre n (*Ce résultat peut être admis pour la suite - on pensera à utiliser des équivalents...*).

2. Pourquoi est-il légitime de penser que la loi des longueurs de segments $Y_{k+1} - Y_k$ peut également être approchée par une loi exponentielle de paramètre n ?

3. Montrer que si $Z \hookrightarrow \exp(\lambda)$ et a est un réel fixé $(a \in \mathbb{R}_+^*)$, alors $T = aZ$ suit encore une loi exponentielle dont on déterminera le paramètre.
   Quelle hypothèse pouvez-vous faire sur la loi de $D_{k+1} - D_k$ où $(D_1, \cdots, D_n)$ désigne les n variables aléatoires ayant pour valeurs les valeurs des variables $(C_1, \cdots, C_n)$ ordonnées dans l'ordre croissant ? (On rappelle que $X_i = C_i/N$)

4. Écrire une fonction Python `simulCassureGenome(N,p)` permettant de retourner une liste `Lc` de $n = 2 * p$ cassures $(C_1, \cdots, C_n)$ uniformément distribué dans un génome de longueur fictive N.

5. Écrire une fonction de tri de votre choix permettant de trier la liste `Lc` pour obtenir une liste `Ld` des mêmes valeurs organisées de façon croissante. En déduire une fonction `longueurSegments(N,p)` qui retourne la liste de toutes les longueurs rencontrées après p inversions.

6. Après avoir répété 100 fois l'opération ci-dessus avec $N = 25000$ et $p = 320$, on obtient des segments de longueur comprise entre 0 et 150. Grâce à la fonction `histogram()` de la bibliothèque `numpy`, on obtient la liste `H` des fréquences relatives des longueurs de segments obtenus. Indiquer un changement de variable et une méthode statistique permettant de montrer que la distribution de ces longueurs suit effectivement une loi exponentielle. Indiquer par ailleurs comment déterminer une estimation de son paramètre λ. On obtient $\lambda = 0.0257$. Cela vous semble-t-il satisfaisant ?

7. On observe ci-dessous l'histogramme des longueurs de séquences homologues supérieures à 1 million de nucléotides (blocs de synthénie) chez la souris et chez l'homme. Pouvez-vous désormais prendre partie dans la querelle qui oppose les tenants des cassures aléatoires et ceux des régions fragiles ?

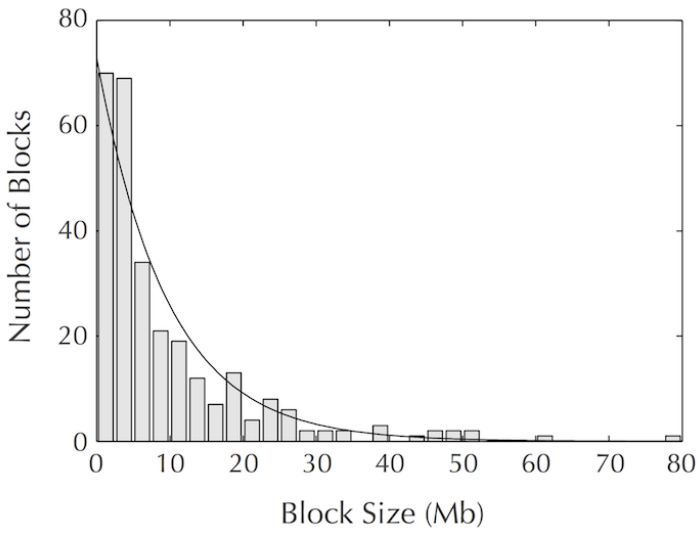

FIGURE 2 – `Histogramme des longueurs de blocs de synthénie homme-souris en 1996, 1416 marqueurs définissant 181 segments conservés (seuls les blocs dont la longueur est supérieure à 1 Mb sont pris en compte). L'histogramme est approché par une distribution exponentielle.`

ANNEXE

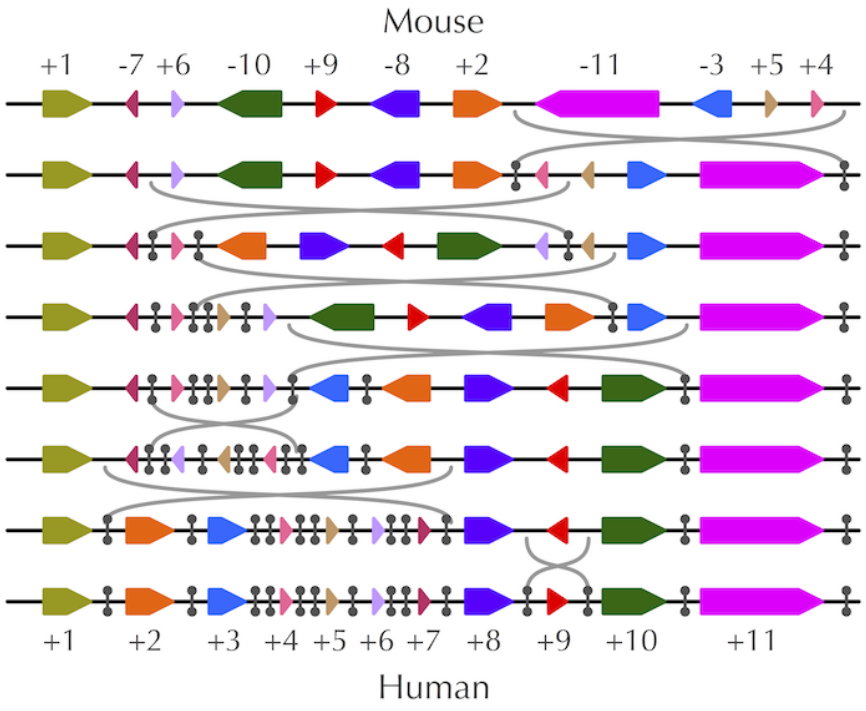

FIGURE 3 – Tranformation du chromosome X de la souris en chromosome X humain avec sept inversions. Chaque bloc synthénique est identifié par une couleur et une forme dont la longueur est proportionnelle à sa taille, associé à un entier compris entre 1 et 11, le signe positif ou négatif de chaque entier indiquant la direction prise par le bloc correspondant. Deux courtes lignes verticales délimitent quant à elles les extrémités du segment inversé au cours de chaque réarrangement.

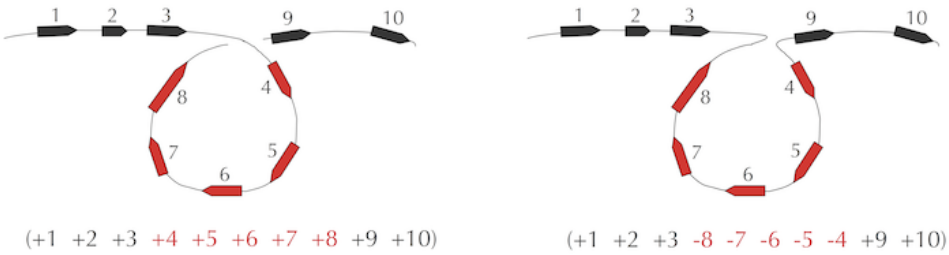

FIGURE 4 – Dessin illustrant comment une inversion casse en deux sites distincts un chromosome et inverse le segment qu'ils délimitent. On notera que l'inversion change le signe de chaque élément au sein du segment permuté.