ADN, dénombrements et probabilités

Problème : Où, dans le génome, débute la réplication de l'ADN ?

Introduction : Nous nous intéressons dans ce problème au processus de réplication dans le cas relativement simple du génome bactérien. Plus précisément nous nous posons la question de savoir comment reconnaître les sites initiant cette réplication au sein du génome.
Nous supposons le chromosome circulaire, comme c'est le cas chez la plupart des procaryotes. Il n'a ni début ni fin et pourtant il contient une séquence d'une centaine de nucléotides (environ 500 dans le cas de *Vibrio cholerae*) appelée « Origine de réplication » au sein de laquelle interviennent les protéines de réplication.
Nous désignerons dans la suite cette séquence par *oriC*.
Connaître l'emplacement d'*oriC* est essentiel, en particulier en thérapie génique. Il est en effet primordial que, dans le cas de rétrovirus, l'intégration dans le génome de la cellule cible du gène artificiel qui code la protéine thérapeutique n'interfère pas avec la duplication de la cellule. Il faut donc, en particulier, éviter comme site d'intégration l'origine de réplication...

Notations et définitions : Une chaîne de caractères désigne une collection ordonnée de symboles sélectionnés dans un alphabet $\mathcal{A}$ dont on peut extraire des sous-chaînes appelées *mot* de longueur le nombre de symboles présents.

- Si $\mathcal{A} = \{0, 1\}$, $S_1 =$« 10111010 » est une chaîne de caractère et $M =$« 011 » est un mot de cette chaîne.
- Si $\mathcal{A} = \{A, C, G, T\}$, la séquence d'ADN $S_2 =$« ATGCTTCAGAAAGGTCTTACG » est une chaîne de caractères de longueur 21. Dans ce contexte, nous appelons **« k-mère »** une sous-séquence ou *mot* de cette chaîne de caractère de longueur k. Par exemple, $M =$« CTTC » est un 4-mère de S_2.

Sous Python, on crée une telle chaîne en écrivant `S2='ATGCTTCAGAAAGGTCTTACG'` qui se manipule ensuite comme une liste.
On accède alors aisément à n'importe lequel des caractères de la chaîne en écrivant `S[i]` pour $i \geq 0$ et à n'importe quelle mot de longueur k (ou k-mère) commençant à la place i (numérotées à partir de 0) en écrivant `S[i:i+k]`. *Exemple :* `S2[0]='A'`, `S2[1]='T'` et `S2[2:6]='GCTT'`

Un mot peut apparaître plusieurs fois dans une séquence d'ADN et on dira qu'un k-mère forme un (L, t)-**bouquet** s'il existe un (court) intervalle du génome de longueur L au sein duquel le k-mère apparaît au moins t fois.

Partie I :

Au sein d'*oriC*, des protéines appelées *DnaA* se lient à certaines sous-séquences, le plus souvent formées de 9 bases (9-mères) et répétées dans *oriC* qu'on appellera par la suite « sites de fixation » ou encore « DnaA box ».
L'ADN, en s'enroulant autour du complexe protéique de *DnaA*, provoquera son ouverture au niveau de ces sous-séquences pour créer deux fourches de réplication. D'un point de vue informatique, ce site s'apparente à un message caché qui indiquerait à *DnaA* « fixe-toi là ! » et en retour la protéine est supposée voir plusieurs sites possibles afin de se fixer plus sûrement...

La fréquence d'apparition de certains 9-mères peut donc être considéré comme un indicateur de site de fixation et, dans cette partie, notre objectif est de mettre en évidence des (L, t)-bouquets au sein d'*OriC*.

Nous travaillerons pour ça sur le chromosome de la bactérie *Vibrio Cholerae* dont le site de réplication est connu. Il est reproduit ci-dessous :

oriC=« ATCAATGATCAACGTAAGCTTCTAAGCATGATCAAGGTGCTCACACAGTTTATCCA
CAACCTGAGTGGATGACATCAAGATAGGTCGTTGTATCTCCTTCCTCTCGTACTCTCATGAC
CACGGAAAGATGATCAAGAGAGGATGATTTCTTGGCCATATCGCAATGAATACTTGTGACTT
GTGCTTCCAATTGACATCTTCAGCGCCATATTGCGCTGGCCAAGGTGACGGAGCGGGATTAC
GAAAGCATGATCATGGCTGTTGTTCTGTTTATCTTGTTTTGACTGAGACTTGTTAGGATAGA
CGGTTTTTCATCACTGACTAGCCAAAGCCTTACTCTGCCTGACATCGACCGTAAATTGATAA
TGAATTTACATGCTTCCGCGACGATTTACCTCTTGATCATCGATCCGATTGAAGATCTTCAA
TTGTTAATTCTCTTGCCTCGACTCATAGCCATGATGAGCTCTTGATCATGTTTCCTTAACCC
TCTATTTTTTACGGAAGAATGATCAAGCTGCTGCTCTTGATCATCGTTTC »

1. Une séquence d'ADN `S` de longueur N étant donnée, ainsi qu'un mot de longueur k, on cherche à savoir si ce mot est présent ou non et, si oui, combien de fois.

   a) Que réalise la ligne de commande : `[S.count(nc)/len(S) for nc in 'ACGT']` ?

   b) Si i désigne la place occupée par la première lettre d'un mot de longueur k au sein de la séquence, dire en fonction de N et de k quelles sont les valeurs possibles prises par i ?

   c) Écrire, après avoir présenté une rapide analyse de votre démarche, une fonction Python `decompte(sequence,mot)` qui retourne le nombre de fois où un k-mère « mot » est présent dans une séquence d'ADN.
   Exemple : `decompte(oriC,'ATGATCAAG')=3` et `decompte(oriC,'TAGATCA')=0`

2. On cherche à obtenir les k-mères les plus fréquents au sein d'OriC, k étant un entier naturel fixé.

   a) Analyser ligne à ligne et compléter si nécessaire la fonction suivante :

```
def effectifMots(sequence,k):
    N=len(sequence)
    compte=[0]*(N-k+1)
    for i in range(...):
        mot=sequence[...:...]
        compte[i] = decompte(sequence, mot)
    return compte
```

   b) Si `sequence='ACAACAATTTGCAATAATTT'` que retourne `effectifMots(sequence,3)` ?

   c) Écrire une fonction `maximum(L)` permettant de retourner le maximum d'une liste `L`.

   d) On dit qu'un k-mère est le plus fréquent si aucun autre k-mère n'est plus fréquent que lui.
   En déduire une fonction `motsLesPlusFrequents(sequence,k)` qui fasse intervenir les fonctions `effectifMots(sequence,k)` et `maximum(L)` et retourne le ou les mots les plus fréquents de longueur k d'une séquence d'ADN donnée ainsi que leur effectif.

Conclusion : En appliquant la fonction précédente à *OriC* de *V. cholerae*, on obtient avec une fréquence égale à 3, les quatre 9-mères suivants :

`ATGATCAAG`, `CTTGATCAT`, `TCTTGATCA`, `CTCTTGATC`

Partie II :

Les quatre 9-mères obtenus précédemment, à cause de leur effectif, sont de bons candidats pour constituer des sites de fixation pour *DnaA*. Pour en décider, il faut pourtant pouvoir dire si cet effectif peut être dû au hasard ou si son caractère exceptionnel doit attirer notre attention.
Cette partie est consacrée à évaluer la probabilité qu'il existe un 9-mère apparaissant trois fois ou plus dans une séquence aléatoire d'ADN de longueur 500.

1. On suppose qu'une séquence S d'ADN de longueur n est un n-uplet d'un alphabet $\mathcal{A} = \{A, C, G, T\}$ de cardinal $a = 4$.
   a) Combien y a-t-il de séquences possibles de longueur n composées avec cet alphabet ?
   b) On considère deux 9-mères de S. Quelle est la probabilité qu'ils possèdent la même première lettre ?
   c) Si M et N sont deux 9-mères formés au hasard à partir de l'alphabet $\mathcal{A}$, quelle est la probabilité que M soit égale à N ?

   Un alphabet de cardinal a étant donné, on cherche à déterminer la probabilité $\mathbb{P}(N, a, mot, t)$ qu'un k-mère *mot* donné apparaisse au moins t fois ($t \in \mathbb{N}$) dans une séquence de longueur N.

2. Nous commençons par un alphabet $\mathcal{A} = \{P, F\}$ de cardinal $a = 2$ en imaginant que chaque lettre correspond au résultat des lancers successifs d'une pièce de monnaie équilibrée.

   a) A titre d'exemple, si S =« PPFP », alors $N = 4$ et S est une séquence au sein de laquelle M_1 =« PF » est un mot de $k = 2$ lettres.
      i. Déterminer combien de séquences de 4 lettres il est possible de former avec un tel alphabet. Décrire explicitement Ω.
      ii. Déterminer la probabilité $\mathbb{P}(4, 2, PF, 1)$ d'obtenir au moins une fois le mot « PF » au sein de cette séquence.
      iii. Déterminer ensuite la probabilité $\mathbb{P}(4, 2, PP, 1)$ d'obtenir au moins une fois le mot PP.

   b) Toujours au sein d'une séquence de $N = 4$ lettres prises dans l'alphabet $\mathcal{A}$ de cardinal $a = 2$, déterminer la probabilité $\mathbb{P}(4, 2, PF, 2)$ que le mot « PF » apparaisse au moins $t = 2$ fois.
   La comparer à la probabilité $\mathbb{P}(4, 2, PP, 2)$ que le mot « PP » apparaisse au moins 2 fois.

   c) On cherche cette fois à déterminer la probabilité $\mathbb{P}(25, 2, PF, 1)$ de voir apparaître dans une séquence de $N = 25$ lettres prises dans l'alphabet $\mathcal{A} = \{P, F\}$ de cardinal 2 le mot PF au moins $t = 1$ fois. Il est désormais inconcevable de décrire explicitement Ω...

   Soit B_k l'événement : « on obtient pour la première fois Pile suivi de Face aux lancers k et $k+1$ ».
      i. Calculer $\mathbb{P}(B_1)$ et $\mathbb{P}(B_2)$
      ii. En considérant un système complet d'événements associé au résultat du premier lancer, montrer que : $\forall k \geq 2$, $\mathbb{P}(B_k) = \dfrac{1}{2}\mathbb{P}(B_{k-1}) + (1/2)^{k+1}$
      iii. Soit la suite $(u_k)_{k\in\mathbb{N}^*}$ définie par $u_k = 2^k\mathbb{P}(B_k)$ pour tout $k \geq 1$. Montrer que la suite (u_k) est une suite arithmétique de raison $1/2$ et de premier terme $u_1 = 1/2$. En déduire l'expression de $\mathbb{P}(B_k)$ pour tout $k \geq 1$.
      iv. Montrer que les B_k, $k \geq 1$ forment un système quasi complet d'événements.

v. Exprimer l'événement : « obtenir le mot PF au moins $t = 1$ fois au sein d'une séquence de $N = 25$ lettres prises dans $\mathcal{A} = \{P, F\}$ » à l'aide des événements B_k. En déduire $\mathbb{P}(25, 2, PF, 1)$ dont on donnera une valeur approchée à 10^{-2} près.

d) On considère cette fois une séquence supposée infinie de lettres P et F et on cherche à comparer la première apparition du mot PF et du mot PP dans cette séquence, en numérotant l'apparition des lettres à partir de 0.

i. Soit R_{PF} variable aléatoire égale au rang du premier mot « PF ».
✐ Attention aux indices. A titre d'exemples :
— Si on obtient la succession F_0, P_1, F_2 alors il faut attendre le troisième lancer (d'indice 2) pour voir apparaître la première fois le mot PF : l'événement $(R_{PF} = 2)$ est réalisé.
— Si on obtient F_0, P_1, P_2, F_3, alors il faut attendre le quatrième lancer (d'indice 3) pour voir apparaître la première fois le mot PF : $(R_{PF} = 3)$ est réalisé.
— Si on obtient P_0, F_1, alors $(R_{PF} = 1)$ est réalisé.

Faire le lien entre l'événement $(R_{PF} = k)$ et l'événement B_k défini en 2.c).
En déduire l'existence de $\mathbb{E}(R_{PF})$ et sa valeur.

ii. Soit R_{PP} variable aléatoire égale au rang du premier mot « PP »
✐ Dans l'exemple 2.d) qui présente la succession de lancers F_0, P_1, P_2, F_3 on a l'événement $(R_{PP} = 2)$ qui est réalisé.

On note $\pi_k = \mathbb{P}(R_{PP} = k)$. Déterminer π_0 et π_1.
Montrer en utilisant la formule des probabilités totales (on pourra admettre ce résultat) que

$$\pi_k = \frac{1}{2}\pi_{k-1} + \frac{1}{4}\pi_{k-2}, \forall k \geq 2$$

En déduire la loi de R_{PP}, à savoir $R_{PP}(\Omega)$ et $\forall k \in R_{PP}(\Omega)$, $\mathbb{P}(R_{PP} = k)$.

iii. Justifier l'existence puis déterminer l'espérance de R_{PP} qu'on comparera à celle de R_{PF}.

Retenons que les rangs d'arrivés des mots dans une séquence dépend de leur composition et que, y-compris les probabilités $\mathbb{P}(N, a, mot, t)$, dépendent du mot considéré. Il est par ailleurs aisé de concevoir que la difficulté pour obtenir cette probabilité augmente avec la longueur de ce mot. Pour cette raison, nous nous contentons dans la suite d'approximer cette probabilité plutôt que de la calculer exactement.

3. Nous supposons cette fois un alphabet $\mathcal{A}$ de cardinal $a = 3$. Considérons par exemple $N = 7$ tirages successifs avec remise dans une urne contenant en égale proportion des boules numérotées 1, 2 et 3. L'alphabet est alors $\mathcal{A} = \{0, 1, 2\}$, $S = 1220110$ est une séquence de longueur $N = 7$ et 01 est un mot de deux lettres.
On cherche à déterminer $\mathbb{P}(7, 3, 01, 2)$

a) Combien y a-t-il de de séquences de 7 lettres possibles ?

b) Montrer qu'il y a exactement $\binom{5}{2}$ tirages possibles amenant deux mots 01 complété par trois 2.

c) En déduire qu'on peut approcher $\mathbb{P}(7, 3, 01, 2)$ par $\dfrac{3^3 \cdot 10}{\text{Card}\Omega}$. Cette approximation vous semble-t-elle bonne ? A t-on obtenu exactement $\mathbb{P}(7, 3, 01, 2)$? une valeur par excès ou une valeur par défaut ?

4. *Généralisation :* Nous cherchons cette fois à approcher $\mathbb{P}(N, a, mot, t)$ probabilité d'obtenir au sein d'une séquence de longueur N formée de lettres prise dans un alphabet $\mathcal{A}$ de cardinal a le k-mère *mot* au mot t fois.

   a) De combien de façon pouvez-vous implanter trois 9-mères (supposés sans recouvrement) dans une séquence d'ADN de longueur 500 ?

   b) Montrer que $\mathbb{P}(N, a, mot, t) \approx p = \dfrac{\binom{N - t(k-1)}{t}}{a^{t \cdot k}}$.

   c) En déduire la probabilité qu'un k-mère quelconque apparaisse au plus t fois ainsi que la probabilité que tous les k-mères apparaissent moins de t fois dans une chaîne aléatoire de longueur N.

   d) Si $\mathbb{P}(N, a, k, t)$ est la probabilité qu'il existe un k-mère apparaissant au moins t fois, montrer que :

$$\mathbb{P}(N, a, k, t) \approx p \cdot a^k = \frac{\binom{N - t(k-1)}{t}}{a^{(t-1) \cdot k}}$$

   e) En déduire que $\mathbb{P}(500, 4, 9, 3) \approx \dfrac{1}{3900}$

   Cette probabilité extrêmement faible de trouver des 9-mères répétés par trois fois dans la région *oriC* de *Vibrio cholerae* nous conduit à poser l'hypothèse suivante : l'un des quatre 9-mères obtenu à l'issue de la partie I représente un potentiel site de fixation pour *DnaA*.

5. Au regard de l'appariement de Watson-Crick ($A - T, C - G$) montrer que deux des 9-mères obtenus sont antiparallèles (on rappelle que si on note p un nucléotide et $\bar{p}$ son complémentaire, alors le complément du mot $p_1 p_2 \cdots p_n$ est obtenu par transcription antiparallèle et vaut : $\bar{p_n} \cdots \bar{p_2} \bar{p_1}$).
   Conclure sur un 9-mère apparaissant six fois au moins dans la région *oriC* (la protéine *DnaA* se fixant aussi bien sur chacun des deux brins complémentaires).
   Cet effectif conclut notre mise en évidence probabiliste du site de fixation.