Devoir maison 3 : Dénombrement et probabilités

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

Partie I :

1. Une séquence d'ADN 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) *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 ?*
   On n'oublie pas qu 'en Python, les lettres de la séquence seront indicées de 0 à $N-1$.
   Tout mot de k lettres à sa première lettre qui peut donc commencer à l'indice $i = 0$ ou, au maximum, à l'indice $i = N - k$ puisque, si le mot est en fin de séquence, sa dernière lettre sera en place $N-1$, son avant-dernière en place $N-2$ et donc sa première en $N-k$.

   Conclusion : $\boxed{i \in [\![0, N-k]\!]}$

   b) *Écrire 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,'ATGATCA')=5` et `decompte(oriC,'TAGATCA')=0`

   Analyse : On commence par déterminer les longueurs N de la séquence et k du mot recherché.
   Après avoir initialisé le compteur à 0, on parcourt la séquence en faisant varier la première place possible i du mot de 0 à $N-k$.
   Le nombre de répétitions étant connues, on utilise pour ça une boucle « pour ».
   Alors, si la sous-séquence de longueur k commençant à la place i est égale au mot cherché, on incrémente le compteur de 1.

   Programme :

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

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

   a) Analyser et expliquer en commentant chacune des lignes le rôle de cette fonction :

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

A la ligne 1 on calcule la longueur de la séquence.
Puis on initialise une liste appelée `compte` formée de $N-k+1$ zéros.
On considère alors successivement tous les mots de k lettres de la séquence, la première lettre prenant successivement les places 0 à $N-k$ et on compte le nombre de fois où ils sont présents.
A l'issue de la fonction, `compte` est une liste contenant l'effectif de chaque mot commençant à la place i.

b) Si `sequence='ACAACAATTTGCAATAATTT'` que retourne `effectifMots(sequence,3)` ?
'ACA' est présent 2 fois donc `compte[0]=2`
'CAA' est présent 3 fois donc `compte[1]=3`
etc.
'TTT' est quant à lui présent 2 fois et donc `compte[17]=2`

Conclusion : `compte = [2, 3, 1, 2, 3, 3, 2, 2, 1, 1, 1, 3, 3, 1, 1, 3, 2, 2]`

c) *Écrivons une fonction* ***`maximum(L)`*** *permettant de retourner le maximum d'une liste* `L` *:*

C'est une fonction bien connue pour laquelle il suffit, par exemple, d'initialiser le maximum au premier terme de la liste puis de parcourir l'ensemble de la liste par une boucle « Pour » afin d'affecter au maximum tout terme de la liste qui lui est supérieur :

```
def maximum(L):
    n=len(L)
    m=L[0]
    for k in range(1,n):
       if L[k]>m:
         m=L[k]
    return m
```

d) On dit qu'un k-mère est le plus fréquent si aucun autre k-mère n'est plus fréquent que lui.
Ecrivons une fonction ***`motsLesPlusFrequents(sequence,k)`*** *qui retourne le ou les mots les plus fréquents de longueur k d'une séquence d'ADN donnée ainsi que leur effectif :*

Analyse : Une idée possible consiste à appeler la fonction `compte=effectifMots(sequence,k)` afin d'obtenir chacun des effectifs des mots de k lettres au sein de la séquence.
On recherche ensuite l'effectif maximum grâce à la fonction `maximum` écrite en c).
Il suffit alors de parcourir toute la liste `compte` et à chaque fois que l'effectif `compte[i]` (pour $i \in [\![0, N-k]\!]$) est maximum, on ajoute le mot commençant à la place i à la liste des mots les plus fréquents.

Programme :

Première version :

```
1  def motsLesPlusFrequents(sequence, k):
2      compte=effectifMots(sequence,k)
3      motifsFrequents = []
4      effmax = max(compte)
5      N=len(sequence)
6      for j in range(N-k+1):
7          if compte[j] == effmax:
8              motifsFrequents.append(sequence[j:j+k])
9
10     return effmax,set(motifsFrequents)
```

✎ la fonction `set` de Python permet d'éviter les répétitions dans une liste.
Ainsi `set([1,4,3,2,4,1,3])={1,2,3,4}`

Deuxième version :

```
1  def motsLesPlusFrequents2(sequence, k):
2      compte=effectifMots(sequence,k)
3      motifsFrequents = []
4      effmax = max(compte)
5      N=len(sequence)
6      for j in range(N-k+1):
7          if compte[j] == effmax and sequence[j:j+k] not in motifsFrequents:
8              motifsFrequents.append(sequence[j:j+k])
9
10     return effmax,motifsFrequents
```

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`, `CTTGATACAT`, `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 $\mathcal{A} = \{A, C, G, T\}$.
   a) *Dénombrons les séquences possibles de longueur n :*
   Il s'agit d'une n-liste d'éléments pris dans un ensemble de cardinal 4 donc :
   $$\boxed{\text{Card}(\Omega) = 4^n}$$
   b) Pour former le début d'un mot, on prélève deux lettres prises au hasard au sein de l'alphabet $\mathcal{A}$. Quelle est la probabilité que ces deux lettres soient les mêmes ?

   Soit M l'événement : « les deux lettres sont identiques ».
   La première lettre de chaque mot formant une 2-liste, on peut répondre à cette question en mettant en évidence les cas favorables, à savoir :
   $$\{(A, A), (C, C), (G, G), (T, T)\}$$

Dès lors, parce que le tirage de chaque lettre est équiprobable, on a :

$$\boxed{\mathbb{P}(M) = \frac{\text{Card}(M)}{\text{Card}(\Omega)} = \frac{4}{16} = \frac{1}{4}}$$

Une deuxième méthode consiste à appliquer la formule des probabilités totales...
On considère le système complet d'événements : $\{A, C, G, T\}$ où A, C, G et T désignent respectivement les événements : « tirer la lettre A, C, G et T ».
Alors, d'après la formule des probabilités totales :

$$\mathbb{P}(M) = \mathbb{P}(M \cap A) + \mathbb{P}(M \cap C) + \mathbb{P}(M \cap G) + \mathbb{P}(M \cap T)$$

Or $\mathbb{P}(M \cap A) = \mathbb{P}(A_1 \cap A_2)$ où A_k désigne l'événement : « tirer la lettre A au k-ième tirage ».
donc $\mathbb{P}(M \cap A) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) = \dfrac{1}{4} \cdot \dfrac{1}{4} = \dfrac{1}{16}$ car les événements sont indépendants.
De même, $\mathbb{P}(M \cap C) = \dfrac{1}{16} = \mathbb{P}(M \cap G) = \mathbb{P}(M \cap T)$
Conclusion : $\boxed{\mathbb{P}(M) = \dfrac{1}{16} + \dfrac{1}{16} + \dfrac{1}{16} + \dfrac{1}{16} = 4 \cdot \dfrac{1}{16} = \dfrac{1}{4}}$

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 ?*
M et N sont égaux si et seulement si les nucléotides qui les composent sont égaux deux à deux.
Posons $M = (M_1, \cdots, M_9)$ et $N = (N_1, \cdots, N_9)$.

$$\mathbb{P}(M = M) = \mathbb{P}\left((M_1 = N_1) \cap \cdots \cap (M_9 = N_9)\right) = \prod_{i=1}^{4} \mathbb{P}(M_i = N_i)$$

car les évènements $(M_i = N_i)$ sont mutuellement indépendants.
Conclusion : $\boxed{\mathbb{P}(M = N) = \left(\dfrac{1}{4}\right)^9}$

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 2 en imaginant des lancers successifs d'une pièce de monnaie équilibrée.

a) Si $N = 4$, $S =$« PPFP » est une séquence 4 lettres au sein de laquelle $M_1 =$« PF » est un mot de $a = 2$ lettres.

i. *Déterminons combien de mots de 4 lettres il est possible de former avec un tel alphabet :*
Chaque mot peut être considéré comme une 4-liste d'un ensemble à 2 éléments.
D'où $\boxed{\text{Card}(\Omega) = 2^4 = 16}$. La description complète est :

$$\begin{aligned}\Omega = \{&(P,P,P,P),(P,P,P,F),(P,P,F,P),(P,P,F,F),(P,F,P,P)\\ &(P,F,P,F),(P,F,F,P),(P,F,F,F),(F,F,P,P),(F,F,P,F)\\ &(F,F,F,P),(F,F,F,F),(F,P,P,P),(F,P,P,F),(F,P,F,P),(F,P,F,F)\}\end{aligned}$$

ii. *Déterminons la probabilité $\mathbb{P}(4, 2, PF, 1)$ d'obtenir au moins une fois le mot « PF » au sein de cette séquence :* Le plus rapide est de passer par l'événement complémentaire, à savoir « ne jamais obtenir le mot « PF ». Or il y a cinq résultats ne faisant pas apparaître le mot « PF » :

$$(P,P,P,P),(F,F,P,P),(F,F,F,P),(F,F,F,F),(F,P,P,P)$$

Dès lors, $\boxed{\mathbb{P}(4,2,PF,1) = 1 - \dfrac{5}{16} = \dfrac{11}{16}}$.

iii. *Déterminons la probabilité* $\mathbb{P}(4,2,PP,1)$ *d'obtenir au moins une fois le mot* PP *:*
Ici il est équivalent de passer ou pas par l'événement complémentaire...
Il y a huit résultats qui font apparaître au moins une fois le mot « PP » :

$$(P,P,P,P),(P,P,P,F),(P,P,F,P),(P,P,F,F),(P,F,P,P),(F,F,P,P),(F,P,P,P),(F,P,P,F)$$

Dès lors, $\boxed{\mathbb{P}(4,2,PP,1) = \dfrac{8}{16} = \dfrac{1}{2} < \mathbb{P}(4,2,PF,1)}$

b) *Toujours au sein d'une séquence de* $N = 4$ *lettres prises dans l'alphabet* $\mathcal{A}$ *de cardinal* 2*, déterminons la probabilité* $\mathbb{P}(4,2,PF,2)$ *que le mot « PF » apparaisse au moins* $t = 2$ *fois :*
Reprenons Ω décrit plus haut et dénombrons les cas favorables.
Il n'y en a qu'un, à savoir ; (P,F,P,F)
Dès lors, $\boxed{\mathbb{P}(4,2,PF,2) = \dfrac{1}{16}}$
De même, il est aisé de voir que

$$\boxed{\mathbb{P}(4,2,PP,2) = \frac{3}{16} > \mathbb{P}(4,2,PF,2)}$$

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.

On pose B_k l'évènement : « on obtient pour la première fois Pile suivi de Face aux lancers k et $k+1$ ».

i. *Calculons* $\mathbb{P}(B_1)$ *et* $\mathbb{P}(B_2)$ *:*
Introduisons les notations suivantes :
Soit P_k l'évènement : « obtenir Pile au k-ième lancer ».
et F_k l'évènement : « obtenir Face au k-ième lancer ».
Les épreuves correspondant aux lancers successifs de la pièce de monnaie étant indépendantes, les évènements associés à ces épreuves sont mutuellement indépendants. D'où on tire :

$$\mathbb{P}(B_1) = \mathbb{P}(P_1 \cap F_2) = \mathbb{P}(P_1) \cdot \mathbb{P}(F_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

et

$$\mathbb{P}(B_2) = \mathbb{P}(P_1 \cap P_2 \cap F_3) + \mathbb{P}(F_1 \cap P_2 \cap F_3) = \mathbb{P}(P_1) \cdot \mathbb{P}(P_2) \cdot \mathbb{P}(F_3) + \mathbb{P}(F_1) \cdot \mathbb{P}(P_2) \cdot \mathbb{P}(F_3) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

ii. *Montrons que :* $\forall k \geq 2$, $\mathbb{P}(B_k) = \dfrac{1}{2}\mathbb{P}(B_{k-1}) + (1/2)^{k+1}$ *:*
On considère le système complet d'événements : $\{P_1, F_1\}$ associé au premier lancer. Alors, d'après la formule des probabilités totales :

$$\mathbb{P}(B_k) = \mathbb{P}(B_k \cap P_1) + \mathbb{P}(B_k \cap F_1) = \mathbb{P}_{P_1}(B_k)\mathbb{P}(P_1) + \mathbb{P}_{F_1}(B_k)\mathbb{P}(F_1)$$

Or, les épreuves étant indépendantes, les événements associés sont mutuellement indépendants et on a :

$$\mathbb{P}_{P_1}(B_k) = \mathbb{P}(P_2 \cap \cdots \cap P_k \cap F_{k+1}) = \mathbb{P}(P_2) \cdots \mathbb{P}(P_k)\mathbb{P}(F_{k+1}) = \frac{1}{2^{k-1}}\frac{1}{2} = \frac{1}{2^k}$$

De plus, si on a obtenu Face au premier lancer, tout se passe comme si on recommençait à attendre l'apparition de la première succession (P,F) au bout de $k-1$ lancers. Donc :

$$\mathbb{P}_{F_1}(B_k) = \mathbb{P}(B_{k-1})$$

Dès lors :

$$\boxed{\mathbb{P}(B_k) = \frac{1}{2^k}\mathbb{P}(P_1) + \mathbb{P}(B_{k-1})\mathbb{P}(F_1) = \frac{1}{2^{k+1}} + \frac{\mathbb{P}(B_{k-1})}{2},\ \forall k \geq 2}$$

iii. Soit la suite $(u_k)_{k\in\mathbb{N}^*}$ définie par $u_k = 2^k\mathbb{P}(B_k)$ pour tout $k \geq 1$. *Montrons que la suite* (u_k) *est une suite arithmétique de raison* $1/2$ *et de premier terme* $u_1 = 1/2$ *:*
D'après ce qui précède, on a :

$$u_1 = 2\mathbb{P}(B_1) = \frac{2}{4} = \frac{1}{2} \text{ et } \forall k \geq 2,\ u_k = 2^k\mathbb{P}(B_k) = 2^{k-1}\mathbb{P}(B_{k-1}) + \frac{1}{2}$$

Conclusion : $\boxed{(u_k) \text{ est une suite arithmétique définie par : } u_1 = \frac{1}{2} \text{ et } \forall k \geq 2,\ u_k = u_{k-1} + \frac{1}{2}}$

On en déduit que $\forall k \geq 1,\ u_k = u_1 + (k-1)\frac{1}{2} = \frac{1+(k-1)}{2} = \frac{k}{2}$.
En divisant par 2^k, on obtient :

$$\boxed{\mathbb{P}(B_k) = \frac{u_k}{2^k} = \frac{k}{2^{k+1}} \text{ pour tout } k \geq 1}$$

iv. *Montrons que les* B_k, $k \geq 1$ *forment un système quasi-complet d'évènements :*
Montrons que la série $\sum \mathbb{P}(B_k)$ converge et que $\sum_{k=1}^{\infty} \mathbb{P}(B_k) = 1$:

$$\sum_{k\geq 1} \mathbb{P}(B_k) = \sum_{k\geq 1} \frac{k}{2^{k+1}} = \sum_{k\geq 1} \frac{1}{4} k (\frac{1}{2})^{k-1}$$

or $\sum_{k\geq 1} k(\frac{1}{2})^{k-1}$ est une série géométrique dérivée convergente car $q \in]0,1[$.

Sa somme vaut $S_1 = \sum_{k=1}^{\infty} k(\frac{1}{2})^{k-1} = \frac{1}{(1-1/2)^2} = 4$

Donc, $\sum_{k\geq 1} \frac{1}{4} k(\frac{1}{2})^{k-1}$ converge car la multiplication par $\lambda = \frac{1}{4}$ ne change par la nature de la série.

Conclusion : $\boxed{\sum_{k\geq 1} \mathbb{P}(B_k) \text{ converge de somme } \sum_{k=1}^{\infty} \mathbb{P}(B_k) = \frac{1}{4}4 = 1}$

Si on ajoute que $\bigcup_{k=1}^{\infty} B_k$ n'est pas Ω puisqu'il s'agit de l'événement « obtenir le mot « PF » lors d'une succession de lancers de pièce de monnaie » alors on a montré qu'il s'agit d'un système quasi-complet d'événements plutôt que de système complet d'événements.

v. *Exprimons 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 *:*
On note qu'on obtient le mot PF au moins une fois au sein d'une séquence de 25 lettres si et seulement si on l'obtient au moins une première fois entre les places 1 et 24.
Dès lors :

$$\mathbb{P}(25, 2, PF, 1) = \mathbb{P}(\bigcup_{k=1}^{24} B_k) = \sum_{k=1}^{24} \mathbb{P}(B_k) = \frac{1}{4} \cdot \sum_{k=1}^{24} k(\frac{1}{2})^{k-1}$$

puisque les événements B_k sont deux à deux incompatibles.

Effectuons le calcul de $T = \sum_{k=1}^{24} k(\frac{1}{2})^{k-1}$:

Soit $P(x) = \sum_{k=0}^{24} x^k = \dfrac{1-x^{25}}{1-x}$

Donc $P'(x) = \sum_{k=0}^{24} kx^{k-1} = \dfrac{-25x^{24}(1-x)+(1-x^{25})}{(1-x)^2}$.

Dès lors, en prenant $x = 1/2$:

$$T = \frac{-\frac{25}{2^{24}}\cdot\frac{1}{2}+1-(\frac{1}{2})^{25}}{\left(\frac{1}{2}\right)^2} = \left(-\frac{25}{2^{25}}+1-\frac{1}{2^{25}}\right)\cdot 4 = \left(1-\frac{26}{2^{25}}\right)\cdot 4$$

Conclusion : $\boxed{\mathbb{P}(25,2,PF,1) = \frac{1}{4}\cdot T = 1-\frac{26}{2^{25}} = 1-\frac{13}{2^{24}} \approx 0.9999}$

✎*Remarque :* Pour obtenir cette réponse, on pouvait se passer d'exprimer $\mathbb{P}(25,2,PF,1)$ en fonction des B_k. En effet, en passant par l'événement contraire :

$$\mathbb{P}(25,2,PF,1) = 1-\mathbb{P}(25,2,PF,0)$$

avec : $\text{Card}(25,2,PF,0) = 26$ si on note que les cas favorables sont :

$$(P_1,\cdots,P_{25}),(F_1,P_2,\cdots,P_{25}),(F_1,F_2,P_3\cdots,P_{25}),\cdots,(F_1,\cdots,F_{24},P_{25}),(F_1,\cdots,F_{25})$$

Dès lors : $\boxed{\mathbb{P}(25,2,PF,1) = 1-\frac{26}{2^{25}} = 1-\frac{13}{2^{25}}}$

On est quasiment sûr d'obtenir au moins une fois le mot PF sur une séquence de 25 lettres prises dans $\{P,F\}$. En est-il de même des 9-mères obtenus à la fin de la partie I alors qu'on considérait des séquences de 500 nucléotides ?

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. Il va de soit que

$$(R_{PF} = k) = B_k$$

où B_k a été défini à la question 2.c).

En effet B_k est réalisé lorsqu'on a obtenu pour la première fois Pile suivi de Face aux lancers k et $k+1$, c'est-à-dire aux lancers d'indices $k-1$ et k, autrement dit lorsque $(R_{PF} = k)$ est réalisé.

Dès lors :

$$\mathbb{E}(R_{PF}) \text{ existe si } \sum k\mathbb{P}(B_k) = \sum \frac{k^2}{2^{k+1}} \text{ converge (absolument).}$$

Je laisse à votre sagacité la preuve de cette convergence en rappelant que $k^2 = k(k-1)+k$. Vous vérifierez à cette occasion qu'on trouve bien :

$$\mathbb{E}(R_{PF}) = \sum_{k=1}^{\infty} k\mathbb{P}(B_k) = \sum_{k=1}^{\infty}\left(\frac{k(k-1)}{2^{k+1}}+\frac{k}{2^{k+1}}\right) = 2+1 = 3$$

ii. Soit R_{PP} variable aléatoire égale au rang du premier mot « PP ».

On note $\pi_k = \mathbb{P}(R_{PP} = k)$.

— *Déterminons π_0 et π_1 :*

π_0 est la probabilité d'obtenir « PP »en un seul lancer. C'est un événement impossible.

Donc $\boxed{\pi_0 = 0}$

$\boxed{\pi_1 = \mathbb{P}(P_1 \cap P_2) = \mathbb{P}(P_1)\cdot\mathbb{P}(P_2) = \frac{1}{4}}$ car les événements sont indépendants.

— *Montrons en utilisant la formule des probabilités totales que :*

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

On considère comme en 2.c) le système complet d'événements : $\{P_1, F_1\}$.
Alors, d'après la formule des probabilités totales :

$$\begin{aligned}\pi_k = \mathbb{P}(R_{PP} = k) &= \mathbb{P}((R_{PP} = k) \cap P_1) + \mathbb{P}((R_{PP} = k) \cap F_1) \\ &= \mathbb{P}_{P_1}(R_{PP} = k) \cdot \mathbb{P}(P_1) + \mathbb{P}_{F_1}(R_{PP} = k) \cdot \mathbb{P}(F_1)\end{aligned}$$

avec $(R_{PP} = k)$ conditionné par P_1 réalisé si, et seulement si, on a Face au deuxième lancer (sinon PP serait réalisé dès le deuxième lancer !) puis PP pour la première fois au bout de $k-2$ lancers, soit

$$(R_{PP} = k) = F_2 \cap (R_{PP} = k-2)$$

et $(R_{PP} = k)$ conditionné par F_1 réalisé si, et seulement si, on obtient pour la première fois PP au bout de $k-1$ lancers.
Dès lors :

$$\begin{aligned}\pi_k = \mathbb{P}(R_{PP} = k) &= \mathbb{P}\left(F_1 \cap (R_{PP} = k-2)\right)\frac{1}{2} + \mathbb{P}(R_{PP} = k-1)\frac{1}{2} \\ &= \mathbb{P}(F_1)\mathbb{P}(R_{PP} = k-2))\frac{1}{2} + \mathbb{P}(R_{PP} = k-1)\frac{1}{2} \\ &\qquad \text{car } F_1 \text{ et } (R_{PP} = k) \text{ indépendants} \\ &= \frac{1}{4}\pi_{k-2} + \frac{1}{2}\pi_{k-1}\end{aligned}$$

Conclusion : $\boxed{(\pi_k) \text{ est une suite récurrente linéaire d'ordre 2}}$

— *déduisons de ce qui précède la loi de R_{PP} :* Exprimons π_k en fonction de k...

Soit (E_c) l'équation caractéristique : $r^2 - \frac{1}{2}r - \frac{1}{4} = 0$

Son discriminant vaut $\Delta = \frac{1}{4} + 1 = \frac{5}{4} > 0$.

(E_c) admet donc deux racines réelles : $r_1 = \frac{1}{2} \cdot \frac{1+\sqrt{5}}{2}$ et $r_2 = \frac{1}{2} \cdot \frac{1-\sqrt{5}}{2}$

et d'après le cours sur les suites récurrentes linéaires d'ordre 2 :

$$\exists a, b \in \mathbb{R},\ \pi_k = a \cdot r_1^k + b \cdot r_2^k,\ \forall k \geq 1$$

Déterminons a et b grâce aux conditions initiales :

$$(S) = \Leftrightarrow \begin{cases} \pi_0 &= a + b = 0 \\ \pi_1 &= ar_1 + br_2 = \frac{1}{4} \end{cases} \Leftrightarrow \begin{cases} b &= -a \\ a(r_1 - r_2) &= \frac{1}{4} \end{cases} \Leftrightarrow \begin{cases} b &= -\frac{\sqrt{5}}{10} \\ a &= \frac{\sqrt{5}}{10} \end{cases}$$

car $r_1 - r_2 = \frac{\sqrt{5}}{2}$

Conclusion : $\boxed{\forall k \geq 0,\ \pi_k = \frac{\sqrt{5}}{10}(r_1^k - r_2^k)}$

✐ Les plus courageux vérifierons que $\pi_k \geq 0$, que la série $\sum \pi_k$ converge et $\sum_{k=0}^{\infty} \pi_k = 1$.

iii. *Justifions l'existence puis déterminons l'espérance de R_{PP} qu'on comparera à celle de R_{PF} :*

Il s'agit de montrer que la série $\sum k\mathbb{P}(R_{PP}=k)$ converge abs. et de calculer sa somme.

Étudions donc la nature de $\sum \frac{\sqrt{5}}{10}(kr_1^k - kr_2^k)$ qui est une série à termes positifs :
On rappelle que $\sum kr_1^{k-1}$ et $\sum kr_2^{k-1}$ convergent en tant que séries géométriques dérivées de raisons respectives r_1 et r_2 toutes deux comprises entre 0 et 1.
Or $k\mathbb{P}(R_{PP}=k) = \frac{r_1\sqrt{5}}{10}kr_1^{k-1} - \frac{r_2\sqrt{5}}{10}kr_2^{k-1}$.

Conclusion : $\boxed{\sum k\mathbb{P}(R_{PP}=k) \text{ converge (abst) par comb. linéaire de séries cvgtes}}$

Alors $\mathbb{E}(R_{PP})$ existe et vaut : $\mathbb{E}(R_{PP}) = \sum_{k=0}^{\infty} k\mathbb{P}(R_{PP}=k)$.

Avant de calculer cette somme, notons au préalable que, puisque r_1 et r_2 sont racines de (E_c) : $r^2 - \frac{1}{2}r - \frac{1}{4} = 0$ on a :

$$r_1 + r_2 = \frac{1}{2} \text{ et } r_1 r_2 = -\frac{1}{4}$$

Dès lors :

$$\begin{aligned}
\sum_{k=0}^{\infty} k\pi_k &= \frac{\sqrt{5}}{10}\left(\frac{r_1}{(1-r_1)^2} - \frac{r_2}{(1-r_2)^2}\right) \\
&= \frac{\sqrt{5}}{10} \cdot \frac{r_1(1-2r_2+r_2^2) - r_2(1-2r_1+r_1^2)}{((1-r_1)(1-r_2))^2} \\
&= \frac{\sqrt{5}}{10} \cdot \frac{r_1 - r_2 + r_1r_2(r_2-r_1)}{((1-r_1)(1-r_2))^2} = \frac{\sqrt{5}}{10} \cdot \frac{(r_1-r_2)(1-r_1r_2)}{((1-r_1)(1-r_2))^2} \\
&= \frac{\sqrt{5}}{10} \cdot \frac{(r_1-r_2)(1-r_1r_2)}{(1-(r_1+r_2)+r_1r_2)^2} \\
&= \frac{\sqrt{5}}{10} \cdot \frac{\sqrt{5}}{2} \cdot \frac{5/4}{(1/4)^2} = \frac{5}{20} \cdot \frac{5}{4} \cdot 16 = 5
\end{aligned}$$

car $r_1 - r_2 = \frac{\sqrt{5}}{2}$ et $1 - r_1r_2 = 1 + \frac{1}{4} = \frac{5}{4}$.

Conclusion : $\boxed{\mathbb{E}(R_{PP}) = 5 > \mathbb{E}(R_{PF}) = 3}$

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 0, 1 et 2. L'alphabet est alors $\mathcal{A} = \{0, 1, 2\}$, $S = 1220110$ est une séquence de longueur N et 01 est un mot de deux lettres.
On cherche à déterminer $\mathbb{P}(7, 3, 01, 2)$

a) Le nombre de séquences de 7 lettres possibles vaut 3^7 puisque chaque séquence est un 7-uplet d'un ensemble $\mathcal{A}$ de cardinal 3.

b) *Montrons qu'il y a exactement* $\binom{5}{2}$ *tirages possibles amenant deux mots* 01 *et un triplet de* 2 *:*
Il y a autant de tirages amenant deux mots 01 dans une séquence de 7 lettres que de façons de placer deux $*$ parmi cinq places possibles.
Par exemple : $(2, *, *, 2, 2)$ ou $(*, *, 2, 2, 2)$ ou encore $(2, 2, 2, *, *)$

Conclusion : Il y a $\binom{5}{2} = 10$ tirages possibles amenant deux 01 et un triplet de 2

c) *Déduisons-en une approximation de* $\mathbb{P}(7, 3, 01, 2)$ *:*
Une fois les deux mots 01 placés, il reste à compléter les trois places libres par un triplet de lettres prises dans $\mathcal{A}$.
Or il y a $3^3 = 27$ 3-uplets possibles d'éléments de $\mathcal{A}$.
Donc, les tirages étant *équiprobables* on peut estimer que :

$$\mathbb{P}(7, 3, 01, 2) \approx \frac{3^3 \cdot 10}{\text{Card}\Omega} = \frac{27 \cdot 10}{3^7} = \frac{270}{2187} \equiv 0.1234567890123456789....$$

(*Ouhaaa... joli nombre !*)

Ce n'est pourtant qu'une approximation car nous n'avons pas obtenu exactement $\mathbb{P}(7, 3, 01, 2)$.
En effet, le nombre de cas favorable est surestimé puisqu'il compte plusieurs fois certaines séquences (en nombre négligeable par rapport à 2187). Par exemple :

— Si le deux mots 01 sont placés en places 0 et 1 : $(*, *, \cdot, \cdot, \cdot) = (\boxed{0,1,0,1}, \cdot, \cdot, \cdot)$ alors les trois « $\cdot$ »étant constitués par tous les triplets possibles de $\mathcal{A}$, on aura en particulier :
$(\boxed{0,1,0,1}, 0, 1, 0)$ ou $(\boxed{0,1,0,1}, 0, 1, 1)$ ou encore $(\boxed{0,1,0,1}, 0, 1, 2)$

— Mots qu'on retrouvera dans le cas où les deux mots 01 sont en places 2 et 3 : $(\cdot, \cdot, *, *, \cdot) = (\cdot, \cdot, \boxed{0,1,0,1}, \cdot)$ alors en complétant les trois points par des triplets de $\mathcal{A}$ on obtiendra en particulier :
$(0, 1, \boxed{0,1,0,1}, 0)$ ou $(0, 1, \boxed{0,1,0,1}, 1)$ ou encore $(0, 1, \boxed{0,1,0,1}, 2)$

Nous pouvons conclure que $\frac{270}{2187}$ est une valeur approchée par excès de $\mathbb{P}(7, 3, 02, 2)$

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 *?*
On généralise ce qui a été fait précédemment. Pour nous aider, nous représentons cette fois encore les trois 9-mère par trois « $*$ » dont il s'agit de trouver les places possibles.
En dehors de ces 9-mères, il reste à choisir $500 - 3 \cdot 9 = 473$ nucléotides et le nombre de places possibles pour les trois « $*$ » est donc égale à $473 + 3 = 476$.
Dès lors, le choix des places étant sans répétition et sans ordre, on a : $\binom{476}{3}$ façons de placer les trois « $*$ » parmi 476 places possibles.

Conclusion : Il y a $\binom{476}{3}$ façons d'implanter les trois 9-mères dans notre séquence.

b) *Montrons que* $\mathbb{P}(N, a, mot, t) \approx p = \dfrac{\binom{N-t(k-1)}{t}}{a^{t\cdot k}}$ *:*
D'après ce qui précède, le nombre de places où implanter nos t k-mères vaut $N - t \cdot k + t$ (en effet, t places pour les t k-mères et $N - t \cdot k$ places pour les autres nucléotides).
Alors, le nombre de places possibles des t k-mères vaut : $\binom{N-t(k-1)}{t}$ Pour chacune de ces places il y a : $a^{N-t\cdot k}$ choix de $N - t \cdot k$ nucléotides choisis dans l'alphabet $\mathcal{A} = \{'A', 'C', 'G', 'T'\}$.
Et comme $\text{Card}(\Omega) = a^N$ (avec ici $a = 4$), on a :

$$\mathbb{P}(N, a, mot, t) \leq \frac{\binom{N-t(k-1)}{t} a^{N-tk}}{a^N}$$

Conclusion : $\boxed{\mathbb{P}(N, a, mot, t) \leq \dfrac{\binom{N-t(k-1)}{t}}{a^{t\cdot k}}}$

c) *Déduisons-en 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 :*
Notons $\dfrac{\binom{N-t(k-1)}{t}}{a^{t\cdot k}}$ la probabilité approximative de $\mathbb{P}(N, a, mot, t)$.
La probabilité qu'un k-mère donné apparaisse au plus t fois est la probabilité de l'évènement contraire et donc cette probabilité vaut $\boxed{1-p}$.

Il y a par ailleurs a^k k-mères possibles formés à partir de cet alphabet.
Chaque k-mère ayant cette même probabilité $1 - p$ d'apparaître au plus t fois, on en déduit par indépendance mutuelle de ces évènements la probabilité que tous les k-mères apparaissent moins de t fois, à savoir :

$$\boxed{(1-p)^{a^k}}$$

d) *On note* $\mathbb{P}(N, a, k, t)$ *la probabilité qu'il existe un k-mère apparaissant au moins t fois.*
D'après ce qui précède :

$$\mathbb{P}(N, a, k, t) = 1 - (1-p)^{a^k} \text{ (évènement contraire)}$$

et comme p est proche de zéro on peut utiliser les équivalents usuels, ici : $(1-u)^\alpha \approx 1 - \alpha u$ et donc :

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

e) On en déduit que

$$\mathbb{P}(500, 4, 9, 3) \approx \frac{\binom{500-3\cdot 8}{3}}{4^{(3-1)\cdot 9}} = \frac{17861900}{68719476736} \approx \frac{1}{3 \cdot 1300} = \frac{1}{3900} \approx 0.00026$$

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 ($A-T, C-G$) il est immédiat que deux des 9-mères obtenus à l'issue de la première partie sont antiparallèles, à savoir $ATGATCAAG$ et $CTTGATACAT$.

La protéine *DnaA* se fixant aussi bien sur chacun des deux brins complémentaires, la probabilité devient quasi nulle qu'un 9-mère apparaissant six fois au moins dans la région *oriC*.
Cette effectif qui ne peut être du au seul hasard permet de conclure notre recherche statistique du site de fixation.