

MATHÉMATIQUES**Durée : 3 heures****L'usage d'une calculatrice est autorisé pour cette épreuve.**

Chaque candidat est responsable de la vérification de son sujet d'épreuve : pagination et impression de chaque page. Ce contrôle doit être fait en début d'épreuve. En cas de doute, le candidat doit alerter au plus tôt le surveillant qui vérifiera et, éventuellement, remplacera le sujet.

Ce sujet comporte 7 pages numérotées de 1 à 7.

Si, au cours de l'épreuve, un candidat repère ce qui lui semble être une erreur d'énoncé, il le signale sur sa copie et poursuit sa composition en expliquant les raisons des initiatives qu'il a été amené à prendre.

Ce sujet est constitué de deux problèmes totalement indépendants.

Problème 1. Distance de Jukes-Cantor

L'ADN est la molécule support de l'information génétique de tout être vivant. En comparant les séquences d'ADN de plusieurs espèces, on espère pouvoir suivre les différentes mutations qui ont amené à ces espèces.

On rappelle que l'ADN est une suite de nucléotides, ces nucléotides étant identifiés par les lettres A, C, G, T.

Repérer uniquement les différences observées dans les séquences d'ADN ne permet pas d'estimer le nombre de mutations subies par l'ADN, car cela ne permet pas d'identifier des mutations en chaîne.

On suppose que la probabilité de mutation d'un nucléotide vers un nucléotide différent fixé, dans la petite période de temps de durée $h \in]0, \frac{1}{3\alpha}[$, est égale à αh , avec $\alpha \in]0, \frac{1}{3}[$ fixé. Par exemple, si un nucléotide est un A, dans la période de temps de durée h , il mutera vers un C avec probabilité αh , vers un G avec probabilité αh et vers un T avec probabilité αh .

Partie A. Un modèle discrétisé

Dans toute cette partie n est un entier naturel et on considère un temps dont la mesure est discrétisée, c'est-à-dire que la mesure du temps s'écoule par à-coups, et non de façon continue ; on prendra arbitrairement $h = 1$.

On définit alors pour tout $n \in \mathbb{N}$ le vecteur $X_n = \begin{pmatrix} \mathbb{P}(A_n) \\ \mathbb{P}(C_n) \\ \mathbb{P}(G_n) \\ \mathbb{P}(T_n) \end{pmatrix}$, où A_n (resp. C_n , G_n , T_n) est l'événement « le nucléotide observé est un A (resp. C, G, T) au bout de n unités de temps ».

- Justifier que si le nucléotide est un A au temps n , alors il restera A au temps $n + 1$ avec une probabilité $1 - 3\alpha$.
- Montrer que pour tout n , on a $X_{n+1} = MX_n$, où M désigne la matrice

$$M = \begin{pmatrix} 1 - 3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1 - 3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1 - 3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1 - 3\alpha \end{pmatrix}.$$

3. Justifier sans calcul que la matrice M est diagonalisable.

4. Justifier que les vecteurs $v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $v_2 = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$, $v_3 = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$ et $v_4 = \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ sont des vecteurs propres de M ,

et donner les valeurs propres associées.

5. Montrer que la famille (v_1, v_2, v_3, v_4) forme une base de $\mathbb{R}^4$, et en déduire les valeurs propres de M , ainsi que la dimension de ses sous-espaces propres.

6. On note P la matrice de $\mathcal{M}_4(\mathbb{R})$ dont les colonnes sont respectivement v_1, v_2, v_3, v_4 . Justifier sans calcul que P est inversible puis donner l'expression des matrices P^{-1} et D telles que $D = P^{-1}MP$.
On pourra calculer P^{-1} à l'aide de la calculatrice.

7. Déterminer, en justifiant, une expression de M^n en fonction de P, P^{-1}, D et n puis l'expression de M^n en fonction de n et α .

On pourra déterminer M^n à l'aide de la calculatrice.

8. Déterminer l'expression de X_n en fonction de n, M et X_0 .

9. On suppose qu'au temps 0, le nucléotide observé est un A. Quelle est la probabilité que le nucléotide observé au bout de n unités de temps soit un A ?

Partie B. Un modèle continu

Dans cette partie, on considérera un temps dont la mesure est continue et pour lequel l'instant initial se situe au temps $t = 0$.

On rappelle alors que dans la période de temps de durée h , avec $h \in]0, \frac{1}{3\alpha}[$, la probabilité qu'un nucléotide X mute vers un nucléotide fixé Y différent de X est égale à αh .

Préliminaires

10. (a) Soit λ un réel strictement positif. Donner l'ensemble des solutions de l'équation différentielle homogène

$$\forall t \in \mathbb{R}, y'(t) = -4\lambda y(t)$$

(b) Rechercher une solution particulière à l'équation

$$\forall t \in \mathbb{R}, y'(t) = \lambda - 4\lambda y(t)$$

et en déduire l'ensemble de ses solutions.

Cette question servira dans la partie B-I.

11. Soit f une fonction strictement positive sur un intervalle. Montrer que f et $\ln \circ f$ ont les mêmes variations sur cet intervalle.

Cette question servira dans la partie B-II.

12. Soit $m \in \mathbb{N}$ et $n \in \mathbb{N}^*$ tels que $m < \frac{3}{4}n$, déterminer l'ensemble des solutions de l'inéquation d'inconnue $d \in \mathbb{R}_+$

$$4m - 3n \left(1 - e^{-\frac{4d}{3}} \right) \geq 0.$$

Cette question servira dans la partie B-II.

B-I.

On s'intéresse pour l'instant à une unique position dans la séquence d'ADN, qui évolue au cours du temps.

On considère alors un nucléotide donné, et on notera pour tout $t \in \mathbb{R}_+$, $p_{zx}(t)$ la probabilité que le nucléotide soit passé de z au temps 0 à x au temps t . Par exemple, $p_{AG}(2, 7)$ est la probabilité que le nucléotide observé passe de A à G en 2, 7 unités de temps.

On définit alors le vecteur $X(t) = \begin{pmatrix} \mathbb{P}(A(t)) \\ \mathbb{P}(C(t)) \\ \mathbb{P}(G(t)) \\ \mathbb{P}(T(t)) \end{pmatrix}$, où $A(t)$ (resp. $C(t)$, $G(t)$, $T(t)$) est l'événement « le nucléotide

observé est un A (resp. C, G, T) au temps t ».

On admet que pour tout $t \in \mathbb{R}_+$ et $h \in]0, \frac{1}{3\alpha}[$, on a :

$$X(t+h) = R(h)X(t),$$

où $R(h)$ désigne la matrice de $\mathcal{M}_4(\mathbb{R})$:

$$R(h) = \begin{pmatrix} 1-3\alpha h & \alpha h & \alpha h & \alpha h \\ \alpha h & 1-3\alpha h & \alpha h & \alpha h \\ \alpha h & \alpha h & 1-3\alpha h & \alpha h \\ \alpha h & \alpha h & \alpha h & 1-3\alpha h \end{pmatrix}.$$

13. Montrer que pour tous temps $t \in \mathbb{R}_+$ et $h \in]0, \frac{1}{3\alpha}[$:

$$\mathbb{P}(A(t+h)) = (1-3\alpha h)\mathbb{P}(A(t)) + \alpha h(1-\mathbb{P}(A(t))).$$

14. On définit la fonction $\varphi_A : t \mapsto \mathbb{P}(A(t))$, et on admet qu'elle est deux fois dérivable sur $\mathbb{R}_+$.

- (a) Calculer le taux d'accroissement $\frac{\varphi_A(t+h)-\varphi_A(t)}{h}$ en fonction de α et $\varphi_A(t)$.
- (b) En déduire que φ_A est dérivable sur $\mathbb{R}_+$ et déterminer l'expression de sa dérivée en fonction de α et de φ_A .

15. En déduire, pour tout réel positif t , la probabilité $\mathbb{P}(A(t))$ en fonction de α , t et $\mathbb{P}(A(0))$.

16. En déduire que pour tout $x \in \{A, T, G, C\}$ et pour tout réel positif t , on a

$$p_{xA}(t) = \begin{cases} \frac{1}{4} (1 + 3e^{-4\alpha t}) & \text{si } x = A \\ \frac{1}{4} (1 - e^{-4\alpha t}) & \text{si } x \neq A \end{cases}.$$

Étant donnée la symétrie du modèle, on admet que ce résultat reste vrai si on remplace A par n'importe quel nucléotide C, G ou T.

B-II.

On considère maintenant une séquence $Z = z_1 z_2 \dots z_n$ d'ADN de longueur n (c'est-à-dire composée de n nucléotides, chacun identifié par sa position i , $1 \leq i \leq n$, dans la chaîne). On admet que chaque nucléotide dans la séquence mute indépendamment des autres.

On suppose que de l'espèce dont provient cette séquence d'ADN Z est issue une autre espèce, dont la séquence d'ADN correspondante est $X = x_1 x_2 \cdots x_n$. On cherche alors à estimer ce qu'on appellera la distance entre ces deux espèces.

On appelle m le nombre de nucléotides différents entre les deux séquences observées Z et X . On suppose que $m < \frac{3}{4}n$; dans le cas contraire, les deux séquences ont trop de différences pour pouvoir être étudiées. On pose alors $d = 3\alpha t$ où t est la variable temps. On appellera *distance de Jukes-Cantor* la valeur de d maximisant la probabilité qu'au bout du temps $t = \frac{d}{3\alpha}$, une mutation aléatoire donne effectivement un nombre m de différences entre les deux chaînes.

Plus précisément, on note $L(d)$ la probabilité que, au temps $t = \frac{d}{3\alpha}$, dans le cas d'une mutation aléatoire, le nombre de différences observé soit effectivement égal à m .

On appellera donc *distance de Jukes-Cantor* la valeur $\hat{d}$ de d qui maximise la fonction L .

17. Justifier que, pour tout réel positif d , $L(d) = \binom{n}{m} q(d)^m (1 - q(d))^{n-m}$, où $q(d) = \frac{3}{4} \left(1 - e^{-\frac{4d}{3}}\right)$.

18. On cherche alors dans cette question à trouver le d maximisant L .

- (a) Calculer sur $\mathbb{R}_+$ la dérivée de $\ln \circ L$, et justifier qu'elle est de même signe que $4m - 3n \left(1 - e^{-\frac{4d}{3}}\right)$.
- (b) En déduire alors le tableau de variations de L sur $\mathbb{R}_+$.
- (c) En déduire que la distance de Jukes-Cantor est donnée sur $\mathbb{R}_+$ par

$$\hat{d} = -\frac{3}{4} \ln \left(1 - \frac{4m}{3n}\right).$$

La distance de Jukes-Cantor permet de donner une estimation probabiliste du nombre de mutations qu'a subi la séquence d'ADN, en prenant en compte les mutations invisibles, par exemple du type A qui mute en C puis en A.

Problème 2. Modèle de Galton-Watson

Dans ce problème, on s'intéresse à une population d'individus, pouvant se reproduire par autofécondation selon une loi de probabilité donnée. Un individu unique de cette population peut donc donner naissance à un ou plusieurs descendants de façon autonome.

On suppose que, pour cette espèce, la reproduction d'un individu est indépendante de celles des autres individus.

On cherche à étudier la probabilité d'extinction de cette espèce.

Partie A. Préliminaires : les fonctions convexes

Dans cette partie, on considère une fonction φ définie sur un intervalle I de $\mathbb{R}$. On suppose que φ est deux fois dérivable sur I .

On dit alors que la fonction φ est *strictement convexe* sur I si la fonction φ' est strictement croissante sur I .

1. Pour un réel $a \in I$, rappeler l'équation de la tangente à la courbe représentative de φ au point d'abscisse a .
2. On veut montrer dans cette question qu'une fonction est strictement convexe sur I si et seulement si sa courbe représentative est strictement au-dessus de ses tangentes sur I , sauf au point de contact avec la tangente.

(a) On suppose φ strictement convexe sur I , et on fixe un réel $a \in I$.

- i. Soit $x > a$. Citer un théorème ou une formule permettant d'affirmer qu'il existe un réel $c \in]a, x[$ tel que :

$$\varphi(x) = \varphi(a) + \varphi'(c)(x - a).$$

On énoncera ses hypothèses.

- ii. En déduire que sur $]a, +\infty[\cap I$, la courbe représentative de φ est strictement au-dessus de la tangente à la courbe en a .
- iii. À l'aide d'un raisonnement similaire pour le cas $x < a$, montrer que la courbe représentative de φ est strictement au-dessus de la tangente à la courbe en a sur $] - \infty, a[\cap I$.

(b) On suppose que la courbe représentative de φ est strictement au-dessus de toutes ses tangentes sur I , sauf aux points de tangences. On fixe a et b dans I , $a < b$.

- i. Montrer que $\varphi'(a) < \frac{\varphi(b) - \varphi(a)}{b - a}$.
- ii. Montrer que $\varphi'(b) > \frac{\varphi(b) - \varphi(a)}{b - a}$.
- iii. En déduire que φ' est strictement croissante sur I , puis que φ est strictement convexe sur I .

Partie B. Un exemple

On considère dans cette partie uniquement que :

- au début de l'année 0, la population est constituée d'un seul individu
- à chaque année $n \in \mathbb{N}$, chaque individu présent donne naissance à 0, 1 ou 2 enfants avec probabilités respectives $q_0 = \frac{1}{8}$, $q_1 = \frac{5}{8}$ et $q_2 = \frac{2}{8}$, puis meurt.

On rappelle que la reproduction des individus est supposée être indépendante.

On notera alors X_n l'événement « au début de l'année n , l'espèce est éteinte », et p_n sa probabilité. On a donc $p_0 = 0$.

3. Calculer le nombre moyen d'enfants par individu dans cette population.
4. Calculer la probabilité p_1 de l'événement X_1 .
5. Justifier que $p_2 = \frac{1}{8} + \frac{5}{8}q_0 + \frac{2}{8}q_0^2$. On pourra éventuellement s'aider d'un arbre de probabilité.
6. De façon générale, montrer que pour tout entier n ,

$$p_{n+1} = f(p_n),$$

où f désigne la fonction définie sur $[0, 1]$ par $f(x) = \frac{1}{8} + \frac{5}{8}x + \frac{2}{8}x^2$.

7. Donner le tableau de variations de la fonction f sur $[0, 1]$, puis en déduire les variations de la suite (p_n) .
8. Montrer que la suite (p_n) converge.
9. Justifier que pour tout $n \in \mathbb{N}$, on a $p_n \leq \frac{1}{2}$.
10. En déduire la limite de (p_n) . En déduire la probabilité que l'espèce ne s'éteigne pas.

Partie C. Le cas général

Dans cette partie, on suppose maintenant que :

- au début de l'année 0, la population est constituée d'un seul individu ;
- à chaque année n appartenant à $\mathbb{N}$, chaque individu présent donne naissance à k appartenant à $\llbracket 0, N \rrbracket$ enfants, puis meurt ;
- $N \in \mathbb{N}^*$ fixé correspond au nombre maximal d'enfants dans cette espèce.

Pour tout $k \in \llbracket 0, N \rrbracket$, on notera $q_k \in [0, 1]$ la probabilité qu'un individu ait exactement k enfants.

On note toujours p_n la probabilité que l'espèce soit éteinte au début de l'année n . On a donc toujours $p_0 = 0$.

11. Quelle est la probabilité d'extinction de l'espèce si $q_0 = 1$? Si $q_0 = 0$?

Dans toute la suite du problème, on suppose donc que $q_0 \in]0, 1[$. On suppose de plus que $q_0 + q_1 < 1$, de sorte qu'il existe un $k \geq 2$ tel que $q_k \neq 0$.

12. On note m le nombre moyen d'enfants par individu. Exprimer m en fonction des q_k , $k \in \llbracket 0, N \rrbracket$.

13. Justifier que pour tout $n \in \mathbb{N}$,

$$p_{n+1} = f(p_n),$$

où f désigne la fonction définie sur $[0, 1]$ par $f(x) = \sum_{k=0}^N q_k x^k$.

14. Justifier que la fonction f est croissante sur $[0, 1]$. En déduire que la suite (p_n) converge.

15. Calculer $f(1)$.

16. Justifier que la fonction f est strictement convexe sur $[0, 1]$, puis donner l'équation de la tangente à la courbe représentative de f en $a = 1$ en fonction de m .

17. Dans cette question, on distingue deux cas selon la valeur de m .

(a) Dans cette question uniquement, on suppose que $m \leq 1$.

- i. Montrer que la tangente à la courbe représentative de f au point d'abscisse 1 est au-dessus de la droite d'équation $y = x$ sur $[0, 1]$.
- ii. En déduire que l'équation $f(x) = x$ admet une unique solution sur $[0, 1]$.
Indication : on pourra utiliser la valeur de $f(1)$ calculée en question 15.
- iii. En déduire la limite de (p_n) .

(b) Dans cette question uniquement, on suppose $m > 1$.

- i. Montrer qu'il existe un réel $c \in]0, 1[$ tel que $f'(c) = 1$.
- ii. En utilisant un raisonnement par l'absurde, montrer que $f(c) \leq c$.
- iii. Montrer alors qu'il existe un réel $\alpha \in]0, c]$ tel que $f(\alpha) = \alpha$.
- iv. Montrer que α est l'unique point fixe de f sur $]0, 1[$.
- v. Montrer alors que la suite (p_n) est à valeurs dans $[0, \alpha]$.
- vi. En déduire la limite de la suite (p_n) .

18. Dans quel(s) cas l'extinction de l'espèce est-elle certaine ?

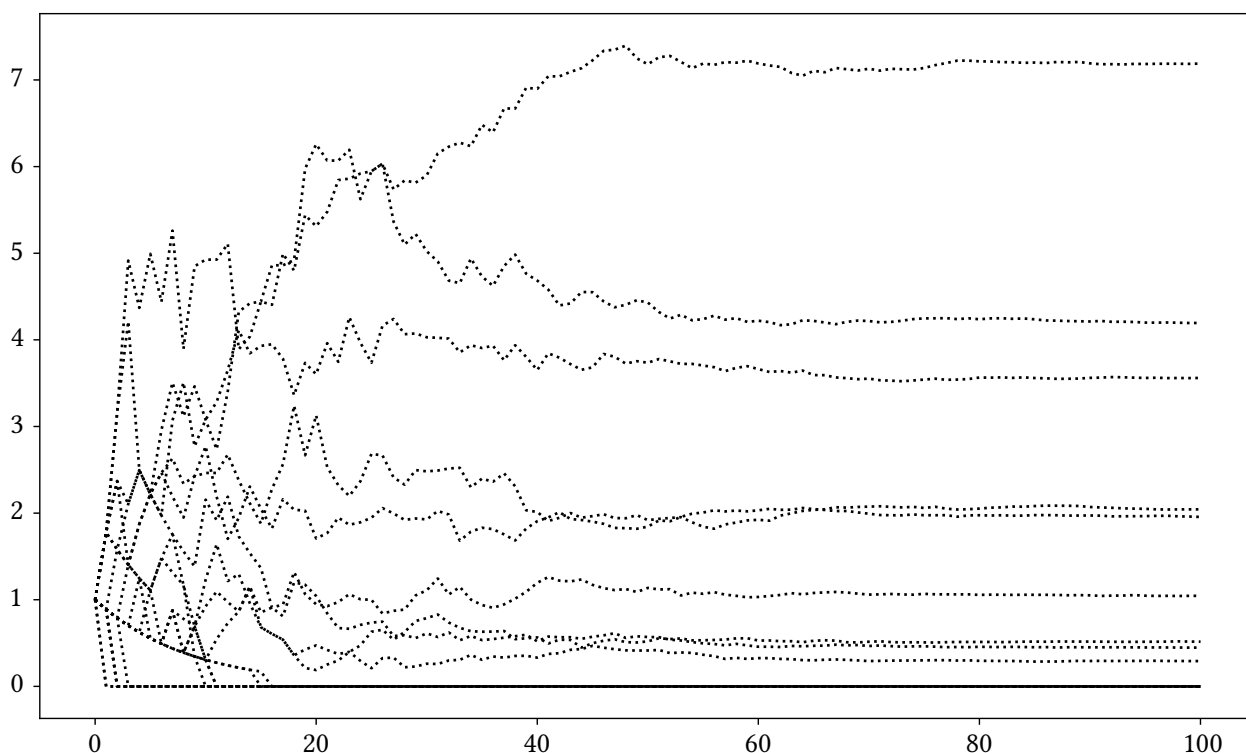
Partie D. Vitesse d'extinction

On se place dans cette partie dans le cas de la partie B, où la loi de reproduction est la suivante : un individu de l'espèce peut avoir

- 0 enfant avec probabilité $\frac{1}{8}$
- 1 enfant avec probabilité $\frac{5}{8}$
- 2 enfants avec probabilité $\frac{2}{8}$

On note dans la suite a_n le nombre d'individus au début de l'année n . On rappelle que m est le nombre moyen d'enfants par individu.

On simule vingt fois l'évolution d'une espèce suivant cette loi de reproduction, et on trace l'évolution de la suite $(\frac{a_n}{m^n})$ en fonction de n . On obtient le graphique suivant :



19. Estimer à partir du graphique la probabilité d'extinction de cette espèce.
20. Dans les cas où l'espèce ne s'éteint pas, que remarque-t-on pour la suite $(\frac{a_n}{m^n})$ au bout d'un temps assez long ?
21. Conjecturer les limites possibles pour la suite (a_n) .