Université Grenoble Alpes **2019-2020**
M1 BÉÉ Biodiversité, écologie, évolution
Modélisation des systèmes écologiques 1

TP – Modèles d'évolution : JC69 vs K80

Format du compte-rendu de TP : **fichier PDF** à envoyer par courriel, **lundi 2 décembre 2019 à 23h59** au plus tard. Titre du message : `[M1BEE] Compte-rendu TP`.

Un compte-rendu rédigé par groupe, un.e ou deux étudiant.e.s par groupe.

Objectifs du TP

Il s'agit d'explorer, en s'aidant de l'outil informatique, le comportement de certains modèles simples d'évolution de l'ADN.

Dans le compte-rendu, on présentera des simulations pour différentes valeurs des paramètres et on commentera les résultats de ces simulations en détail. Le cas échéant, on analysera également les différences entre le comportement théorique des modèles, prédit par leur analyse mathématique, et le comportement observé dans les simulations.

On demande d'inclure *le code en clair d'au moins un des programmes écrits pour les simulations*, par exemple le plus élaboré.

À condition de respecter ces contraintes et de répondre aux questions énumérées ci-dessous, le format du compte-rendu est libre, au sens où on peut y inclure les prolongements de son choix : les leçons à tirer de ces modèles, leur pertinence biologique, leurs limitations, etc. De tels prolongements, à condition d'être pertinents, sont en règle générale très appréciés.

Par contre, la reproduction, au mot près ou quasiment, de pages web ou d'autres sources sur le sujet, sera jugée à la mesure de l'effort que cette opération aura demandé, c'est-à-dire comme étant d'une valeur à peu près nulle. On rappelle par ailleurs que ces procédés sont en général détectables.

Le sujet comprend deux parties : Première partie : Modèles de Jukes et Cantor ; Seconde partie : Modèles de Kimura.

Première partie : Modèles de Jukes et Cantor

Conçus par Thomas H. Jukes et Charles R. Cantor en 1966 et publiés trois ans plus tard [1], ces modèles fournissent la modélisation la plus simple possible de l'évolution d'une séquence non codante d'ADN, sans pression sélective, sous le seul effet de substitutions aléatoires.

En temps discret, pour chaque entier $t \geqslant 0$, la séquence d'ADN au temps t est représentée par une suite $x_t = (x_t^n)_{1 \leqslant n \leqslant N}$ de longueur N donnée. Chaque x_t^n est un élément de l'alphabet nucléotidique $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, et x_t^n représente le nucléotide qui occupe le site n après t intervalles de temps.

Dans le modèle de Jukes et Cantor de paramètre p, pour tous t et n, $x_{t+1}^n = \mathtt{z}$ avec probabilité p, pour chacun des trois éléments $\mathtt{z}$ de $\mathcal{A} \setminus \{x_t^n\}$ (substitution de x_t^n par $\mathtt{z}$), sinon $x_{t+1}^n = x_t^n$ (pas de substitution au temps t au site n), le tout indépendamment de l'historique du site n et indépendamment du comportement des autres sites. Le taux total de substitution en chaque site vaut donc $3p$ et, pour que le modèle ne soit pas absurde, on suppose que $0 < p < \frac{1}{3}$.

1. On fixe $x_0^1 = \mathtt{a}$. Calculer la probabilité d'observer $(x_1^1, x_2^1, x_3^1, x_4^1, x_5^1) = (\mathtt{a}, \mathtt{a}, \mathtt{g}, \mathtt{c}, \mathtt{c})$.

2. Pour tout $t \geqslant 0$, on note m_t la probabilité d'un mismatch $x_t^1 \neq x_0^1$ en position 1 au temps t entre la séquence x_t et la séquence initiale x_0.

Exprimer m_{t+1} en fonction de m_t et p. En déduire la valeur de m_t pour tout $t \geqslant 0$.

3. On admet que, pour des séquences assez longues, la proportion empirique de mismatches entre x_0 et x_t, définie comme $m(x_0, x_t) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\{x_t^n \neq x_0^n\}$, se concentre autour de sa valeur théorique m_t.

En déduire un estimateur $\vartheta(u, v)$ du temps t écoulé entre deux séquences $u = (u^n)_{1 \leqslant n \leqslant N}$ et $v = (v^n)_{1 \leqslant n \leqslant N}$ de longueur N, basé sur la proportion $m(u, v)$ de leurs mismatches.

4. Simuler l'évolution d'une séquence d'ADN de longueur N sous l'effet des substitutions du modèle de Jukes et Cantor de paramètre p pendant T étapes de temps, pour plusieurs ensembles de valeurs (assez petites) du paramètre p et (assez grandes) des longueurs N et T de la séquence d'ADN et de l'intervalle de temps considéré.

1. T. H. Jukes, C. R. Cantor (1969). Evolution of protein molecules. In : H. N. Munro, editor, Mammalian Protein Metabolism, Academic Press, New York, 21-132.

Pour chaque triplet (p, N, T) choisi, il s'agit donc d'obtenir des tableaux $(x_t^n)_{0\leqslant t\leqslant T, 1\leqslant n\leqslant N}$ de taille $(T+1)\times N$ à valeurs dans l'alphabet $\mathcal{A}$.

5. Pour chaque simulation obtenue, calculer $(m(x_0, x_t))_{0\leqslant t\leqslant T}$ le vecteur des proportions successives de mismatches avec la séquence initiale x_0.

Représenter sur un même diagramme les graphes des fonctions $t \mapsto m(x_0, x_t)$ et $t \mapsto m_t$. Commenter.

6. Pour chaque simulation obtenue, calculer le vecteur $(\vartheta(x_0, x_t))_{0\leqslant t\leqslant T}$ et représenter le graphe de la fonction $t \mapsto \vartheta(x_0, x_t)$. Commenter.

Seconde partie : Modèles de Kimura

On souhaite à présent comparer les prédictions basées sur les modèles de Jukes et Cantor et celles basées sur les modèles d'évolution les plus simples après ceux de Jukes et Cantor, introduits par Motoo Kimura en 1980 [2]. L'objectif des modèles de Kimura est de rendre compte de la disparité entre les taux de transition et les taux de transversion. En effet, Kimura remarque que, par exemple dans le génome humain, on observe de 2 à 3 fois plus de transitions que de transversions.

On rappelle que les bases `a` et `g` (adénine et guanine) sont des purines alors que `c` et `t` (cytosine et thymine) sont des pyrimidines, qu'une transition correspond à la substitution d'une purine par une autre purine ou d'une pyrimidine par une autre pyrimidine, et qu'une transversion correspond à la substitution d'une purine par une pyrimidine, ou inversement.

7. Calculer le rapport théorique du nombre de transitions au nombre de transversions sous un modèle de Jukes et Cantor.

8. Dans un modèle de Kimura, on suppose que chaque transition possible en chaque site se produit avec probabilité p_{tr} et que chaque tranversion possible en chaque site se produit avec probabilité p_{tv}. Calculer le taux total de substitution en chaque site. Préciser pour quelles valeurs de p_{tr} et p_{tv}, on retombe sur un modèle de Jukes et Cantor. Enfin, préciser pour quels ratios $p_{\mathrm{tr}}/p_{\mathrm{tv}}$, on obtient un des ratios observés par Kimura, par exemple le ratio de 3 transitions pour 1 transversion.

9. On souhaite désormais comparer les prédictions du modèle de Kimura de paramètres $(p_{\mathrm{tr}}, p_{\mathrm{tv}})$ à celles du modèle de Jukes et Cantor de même taux total de substitution.

2. M. Kimura (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16 (2) : 111-120.

Calculer la valeur p du paramètre de Jukes et Cantor qui correspond au même nombre total de substitutions que les taux de substitutions $(p_{\mathrm{tr}}, p_{\mathrm{tv}})$ dans un modèle de Kimura.

10. Simuler l'évolution d'une séquence d'ADN sous l'effet des substitutions du modèle de Kimura de paramètres $(p_{\mathrm{tr}}, p_{\mathrm{tv}})$, pour plusieurs ensembles de valeurs (assez petites) des paramètres de substitution $(p_{\mathrm{tr}}, p_{\mathrm{tv}})$ et (assez grandes) de N et T.

Pour chaque quadruplet $(p_{\mathrm{tr}}, p_{\mathrm{tv}}, N, T)$ choisi, il s'agit donc d'obtenir des tableaux $(y_t^n)_{0\leqslant t\leqslant T, 1\leqslant n\leqslant N}$ de taille $(T+1)\times N$ à valeurs dans l'alphabet $\mathcal{A}$.

11. Pour chaque simulation $(y_t)_{0\leqslant t\leqslant T}$ obtenue à la question 10, comparer la vitesse de la divergence observée entre les séquences $y_0 = (y_0^n)_{1\leqslant n\leqslant N}$ et $y_t = (y_t^n)_{1\leqslant n\leqslant N}$, à la vitesse de la divergence observée entre les séquences x_0 et x_t simulées dans la première partie, évoluant selon le modèle de Jukes et Cantor de paramètre p calculé à la question 9.

En déduire le comportement de l'estimation de t par l'estimateur $\vartheta(y_0, y_t)$ basé sur le modèle de Jukes et Cantor mais appliqué à des séquences évoluant selon un modèle de Kimura. Commenter et expliquer.