
VERS UNE PHYLOGÉNIE POUR DES MODÈLES D'ÉVOLUTION AVEC INFLUENCE DU VOISINAGE

par

Mikael Falconnet

Stage de M2R 2006-2007

Directeur de stage : Didier Piau

Résumé. — Le récent article de Bérard, Gouéré et Piau sur les processus de substitution avec influence du voisinage fournit des résultats que nous exploitons dans ce papier. Après un rappel sur les modèles classiques d'évolution, nous étudions les caractéristiques des modèles avec influence du voisinage, et nous montrons par exemple que certaines quantités convergent plus rapidement vers leur valeur limite que dans le modèle d'évolution avec indépendance correspondant au même nombre global moyen de mutations par unité de temps, alors que d'autres convergent moins rapidement. Notre but à terme est d'être capable de fournir une phylogénie pour des séquences actuelles et cela demande d'être capable de donner un estimateur du temps écoulé entre une séquence ancestrale et une séquence actuelle pour les modèles d'évolution avec influence du voisinage.

Table des matières

Introduction	1
Partie I. Modèles d'évolution d'une séquence d'ADN	2
1. Premiers modèles.....	3
2. Modèles avec influence du voisinage.....	6
3. Évolutions des modèles.....	9
Partie II. Arbre phylogénétique ou dendrogramme	18
4. Quelques rappels sur la théorie des graphes.....	18
5. Arbres et X-arbres.....	19
Partie III. Reconstruction de l'arbre phylogénétique de séquences d'ADN	21
6. Les mesures de proximité et leurs représentations.....	22
Conclusion	24
Références.....	25

INTRODUCTION

Notre travail comporte deux aspects. Le premier consiste à rendre compte des différents modèles probabilistes utilisés pour modéliser le processus d'évolution d'une

séquence d'ADN, en particulier le modèle de Jukes et Cantor et le modèle de Kimura, qui sont des modèles où l'on suppose que les sites évoluent indépendamment les uns des autres, et les modèles avec influence du voisinage qui nous intéressent plus particulièrement.

Une séquence de nucléotides est répliquée au cours du temps et chaque réplication peut entraîner des mutations de la séquence. Les instants, les sites et la nature des mutations sont les phénomènes qui seront aléatoires dans notre étude. Nous expliquerons les choix faits pour modéliser les mutations et notre but est d'estimer suivant le modèle d'évolution choisi, la distance qui sépare deux séquences d'ADN dont on sait que l'une est issue de l'autre, le terme distance étant à préciser, mais cela peut être par exemple le temps qui s'est écoulé entre ces deux séquences. Quand nous disposerons de cette estimation, notre but sera de l'utiliser pour estimer la distance qui sépare deux séquences actuelles issues d'une même séquence originelle.

Le deuxième aspect ne comporte pas d'aléatoire mais l'utilise, notre objectif est de définir la notion d'arbre phylogénétique et d'expliquer la construction de l'arbre phylogénétique de plusieurs séquences d'ADN que l'on sait être issues d'une même séquence originelle, quand on connaît les distances temporelles qui les séparent. Reconstruire l'arbre phylogénétique de plusieurs séquences d'ADN soulève par exemple des problèmes d'unicité et de robustesse.

PARTIE I MODÈLES D'ÉVOLUTION D'UNE SÉQUENCE D'ADN

Dans les modèles d'évolution de séquences de nucléotides simples, on suppose que chaque site de la séquence d'ADN évolue indépendamment des autres sites et suivant un noyau markovien avec des taux de substitution spécifiques. Nous introduisons des notations couramment utilisées par les biologistes et de plus spécifiques qui s'avèrent pratiques pour décrire les modèles.

Définition 0.1. — *L'alphabet des nucléotides est*

$$\mathcal{A} = \{A, T, C, G\}.$$

Ce sont les premières lettres respectives pour Adénine, Thymine, Cytosine et Guanine. Les nucléotides A et G sont des purines représentées par R, les nucléotides T et C des pyrimidines, représentées par Y. Les substitutions du type $R \leftrightarrow R$ et $Y \leftrightarrow Y$ sont appelées transitions, celles du type $R \leftrightarrow Y$ sont appelées transversions.

On distingue les transitions des transversions car expérimentalement, il arrive dans beaucoup de cas que les transitions soient plus fréquentes que les transversions (pour plus de détails voir [4]). De même, les mesures effectuées sur diverses séquences d'ADN contredisent l'hypothèse d'absence d'interaction entre les sites, d'où la nécessité d'étudier les modèles avec influence du voisinage en s'inspirant des modèles simples, ce qui est l'objet de cette partie.

D'autres notations sont nécessaires pour la suite.

Définition 0.2. — *Pour tout $N_1, N_2, N_3 \in \mathcal{A}$, pour tout $t, s \geq 0$, on désignera par*

- \bar{N}_1 le complémentaire de $\{N_1\}$ dans l'ensemble \mathcal{A} ,
- $F(N_1)(t)$ la fréquence des sites occupés par un nucléotide de type N_1 au temps t ,

- $F(N_1, N_2)(t)$ la fréquence des sites qui étaient occupés au temps 0 par un nucléotide de type N_1 et qui sont occupés au temps t par un nucléotide de type N_2 ,
- $F(N_1, N_2, N_3)(t, s)$ la fréquence des sites occupés par un nucléotide de type N_1 au temps 0, occupés par un nucléotide de type N_2 au temps t et occupés par un nucléotide de type N_3 au temps $t + s$,
- $F(N_2|N_1)(t)$ la probabilité qu'au temps t un site soit occupé par le nucléotide N_2 sachant qu'il était occupé au temps 0 par le nucléotide N_1 .

Les notations précédentes s'adaptent aux dinucléotides. On notera de plus, $*N_1$ respectivement N_1^* l'ensemble des dinucléotides terminant par le nucléotide N_1 , respectivement commençant par le nucléotide N_1 .

Par exemple, $F(C^*, CG)(t)$ désigne la fréquence des 2-sites occupés au temps 0 par des dinucléotides commençant par C et qui sont occupés au temps t par des dinucléotides CG .

1. Premiers modèles

1.1. Modèle de Jukes et Cantor. — Comme nous l'avons dit en introduction, la séquence d'ADN est répliquée au cours du temps, ces phénomènes sont discrets, mais les mutations sont des événements rares, donc en adaptant l'échelle de temps, nous avons recours à une modélisation continue de l'évolution, sous forme de processus à sauts. Une mutation a lieu quand un site de la séquence change d'état. En chaque site, le processus de substitutions est modélisé par un processus de Markov où les états sont les éléments de \mathcal{A} .

1.1.1. Matrice de transition. — Dans le modèle de Jukes et Cantor, souvent noté JC69, on suppose que les sites évoluent indépendamment les uns des autres et que chaque substitution a le même taux λ d'occurrence. En particulier, le taux moyen de substitution par unité de temps est 3λ . Notons Q_{JC} le générateur infinitésimal du processus de Markov de chaque site, on a :

$$Q_{JC} = \begin{matrix} & A & T & C & G \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{matrix}.$$

On peut déterminer la matrice de transition du site en fonction du temps t . Il vient

Proposition 1.1. — Pour tout $t \geq 0$, la matrice de transition au temps t issue du générateur infinitésimal Q_{JC} est

$$P_t = \begin{matrix} & A & T & C & G \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} 1 - 3p(t) & p(t) & p(t) & p(t) \\ p(t) & 1 - 3p(t) & p(t) & p(t) \\ p(t) & p(t) & 1 - 3p(t) & p(t) \\ p(t) & p(t) & p(t) & 1 - 3p(t) \end{pmatrix} \end{matrix}$$

avec

$$p(t) = \frac{1 - e^{-4\lambda t}}{4}.$$

Démonstration. — En introduisant la matrice $J = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$, il vient

$$Q_{JC} = \lambda(-4I + J),$$

avec I la matrice identité qui commute avec J . Or $J^2 = 4J$, ce qui entraîne

$$e^{\lambda t J} = \sum_{n \geq 0} (\lambda t J)^n = I + \frac{J}{4} (e^{4\lambda t} - 1).$$

On termine en multipliant $e^{\lambda t J}$ par $e^{-4\lambda t I}$. \square

1.1.2. Fréquences stationnaires. — Nous pouvons déterminer les fréquences stationnaires pour ce modèle. Etant donné que pour tout $t \geq 0$, P_t est une matrice bistochastique, la loi uniforme est stationnaire pour P_t . Pour chaque nucléotide, la fréquence stationnaire est donc $\frac{1}{4}$.

1.1.3. Estimateur du temps écoulé. — Si l'on appelle q la probabilité qu'un site au temps t soit différent du site initial, on a avec $d = 3\lambda t$ qui correspond à l'échelle de temps liée au taux moyen de substitution

$$q = \frac{3}{4}(1 - e^{-4d/3})$$

Déterminons un estimateur \hat{d} par la méthode du maximum de vraisemblance pour d . Soit x le nombre de sites différents entre les deux séquences et L la longueur de la séquence. Comme les sites se comportent indépendamment, x correspond au nombre de piles obtenus au cours de L lancers de pile ou face avec probabilité q d'obtenir pile. On a donc la vraisemblance

$$f(x, d) = \binom{L}{x} q^x (1 - q)^{L-x} = \binom{L}{x} \left(\frac{3}{4} - \frac{3}{4} e^{-4d/3} \right)^x \left(\frac{1}{4} + \frac{3}{4} e^{-4d/3} \right)^{L-x}.$$

Maximiser $f(x, d)$ en fonction de d est équivalent à maximiser la fonction

$$g(x, d) = x \log \left(\frac{3}{4} - \frac{3}{4} e^{-4d/3} \right) + (L - x) \log \left(\frac{1}{4} + \frac{3}{4} e^{-4d/3} \right)$$

Or elle est maximum en $-\frac{3}{4} \log \left(1 - \frac{4x}{3L} \right)$. On en déduit un estimateur du temps écoulé.

Proposition 1.2. — *Dans le modèle de Jukes et Cantor, un estimateur du temps écoulé est :*

$$\hat{d} = -\frac{3}{4} \log \left(1 - \frac{4}{3} \cdot \frac{x}{L} \right).$$

1.2. Le modèle de Kimura. — Dans le modèle de Kimura, souvent noté K80, on distingue les transitions ($T \leftrightarrow C$ et $A \leftrightarrow G$) des transversions (T ou $C \leftrightarrow A$ ou G), car comme nous l'avons rapidement évoqué précédemment, dans la réalité les transitions sont plus fréquentes que les transversions.

1.2.1. Matrice de transition. — Le taux associé à une transition est α , celui associé à une transversion est β . Le taux moyen de substitution par unité de temps est $\alpha + 2\beta$. Le générateur infinitésimal associé à ce processus de Markov est :

$$Q_K = \begin{matrix} & \begin{matrix} A & T & C & G \end{matrix} \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} -(\alpha + 2\beta) & \beta & \beta & \alpha \\ \beta & -(\alpha + 2\beta) & \alpha & \beta \\ \beta & \alpha & -(\alpha + 2\beta) & \beta \\ \alpha & \beta & \beta & -(\alpha + 2\beta) \end{pmatrix} \end{matrix}$$

Proposition 1.3. — Pour tout $t \geq 0$, la matrice de transition au temps t issue du générateur infinitésimal Q_K est

$$e^{Q_K t} = \begin{matrix} & A & T & C & G \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_2(t) \\ p_1(t) & p_0(t) & p_2(t) & p_1(t) \\ p_1(t) & p_2(t) & p_0(t) & p_1(t) \\ p_2(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix} \end{matrix}$$

avec

$$\begin{aligned} p_0(t) &= \frac{1}{4} \left(1 + e^{-4\beta t} + 2e^{-2(\alpha+\beta)t} \right) \\ p_1(t) &= \frac{1}{4} \left(1 - e^{-4\beta t} \right) \\ p_2(t) &= \frac{1}{4} \left(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t} \right) \end{aligned}$$

Démonstration. — En notant $J_1 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$ et $J_2 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$, on a

$$Q_2 = -2(\alpha + \beta)I + \alpha J_1 + \beta J_2.$$

Or J_1 et J_2 vérifient $J = J_1 + J_2$, $J_1 J_2 = J_2 J_1 = 2J_2$ et $J_1^2 = 2J_1$. Ces propriétés permettent d'établir que

$$e^{\alpha t J_1} = I + \frac{J}{2}(e^{2\alpha t} - 1) \quad \text{et} \quad e^{\beta t J_2} = e^{\beta t J} e^{-\beta t J_1} = e^{-\beta t J_1} e^{\beta t J}.$$

□

1.2.2. Fréquences stationnaires. — De même que pour le modèle JC69, la fréquence stationnaire de chaque nucléotide est $\frac{1}{4}$.

1.2.3. Estimateur du temps écoulé. — Dans ce modèle, notre estimateur doit tenir compte des proportions de transitions S et de transversions V entre les deux séquences. La fonction log-vraisemblance est donnée par

$$f(S, V, \alpha, \beta) = n(1 - S - V) \log\left(\frac{p_0}{4}\right) + nS \log\left(\frac{p_1}{4}\right) + nV \log\left(\frac{p_2}{4}\right).$$

Il en sort

$$\widehat{\alpha t} = -\frac{1}{2} \log(1 - 2S - V) + \frac{1}{4} \log(1 - 2V),$$

et

$$\widehat{\beta t} = -\frac{1}{4} \log(1 - 2V),$$

ce qui donne finalement

$$\hat{d} = -\frac{1}{2} \log(1 - 2S - V) - \frac{1}{4} \log(1 - 2V).$$

1.3. Modèle de Rzhetsky et Nei avec respect de la complémentarité des nucléotides. — Dans le modèle de Rzhetsky et Nei symétrique, en plus de distinguer les transitions des transversions, nous distinguons également le type de nucléotide que la substitution produit. Toutefois, nous respectons la complémentarité des nucléotides dans la double hélice.

1.3.1. *Matrice de transition.* — Le générateur infinitésimal associé à ce modèle est :

$$Q_{W/S} = \begin{matrix} & A & T & C & G \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} \cdot & v_W & v_S & w_S \\ v_W & \cdot & w_S & v_S \\ v_W & w_W & \cdot & v_S \\ w_W & v_W & v_S & \cdot \end{pmatrix} \end{matrix}$$

où les coefficients diagonaux sont tels que la somme sur chaque ligne de $Q_{W/S}$ est nulle.

1.3.2. *Fréquences stationnaires.* — Les fréquences stationnaires pour ce modèle sont différentes de celles pour les modèles JC69 et K80. Le noyau de $Q_{W/S}^T$ est engendré par le vecteur $(v_W + w_W, v_W + w_W, v_S + w_S, v_S + w_S)^T$. On en déduit donc les fréquences stationnaires pour ce modèle.

Proposition 1.4. — *Les fréquences stationnaires pour le modèle de Rzhetsky et Nei symétrique sont :*

$$F(A)(0) = F(T)(0) = \frac{v_W + w_W}{2(v_W + w_W + v_S + w_S)},$$

$$F(C)(0) = F(G)(0) = \frac{v_S + w_S}{2(v_W + w_W + v_S + w_S)}.$$

1.4. **Modèle de Rzhetsky et Nei général.** — En cours d'étude

$$Q_{R/Y} = \begin{matrix} & A & T & C & G \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} \cdot & v_T & v_C & w_G \\ v_A & \cdot & w_C & v_G \\ v_A & w_T & \cdot & v_G \\ w_A & v_T & v_C & \cdot \end{pmatrix} \end{matrix}$$

2. Modèles avec influence du voisinage

Dans les modèles précédents, il n'y a pas d'interaction entre les sites, chaque nucléotide évolue suivant le même processus de Markov et converge en loi vers la distribution stationnaire associée à la matrice des taux. Proche de l'état d'équilibre, on a alors pour tous nucléotides x et y , $F(xy) = F(x)F(y)$, où $F(x)$ est la fréquence du nucléotide x et $F(xy)$ la fréquence du dinucléotide xy . Ceci n'est pas conforme à la réalité. Par exemple, la fréquence du dinucléotide CG dans le génôme humain est cinq fois plus petite que le produit des fréquences des nucléotides C et G (voir [3]).

Les modèles que nous allons étudier à présent prennent en compte la nature des sites voisins dans l'évolution d'un site, en particulier nous allons introduire ce que nous appellerons des mutations doubles pour illustrer l'instabilité des îlots CG . Cette approche plus réaliste a des conséquences sur l'estimation de la distance entre deux séquences d'ADN, par exemple connaître la fréquence $F(x)$ du nucléotide x nécessite de connaître celle des trinuécléotides $F(yxz)$, etc.

2.1. **Modèle 1+r.** — Le modèle 1+r est le modèle non trivial le plus simple qui prend en compte cette particularité de l'influence des voisins. Nous allons perturber le modèle de Jukes et Cantor en donnant une possibilité supplémentaire de mutation aux îlots CG . Le taux de mutation pour les nucléotides impliqués dans les îlots CG est augmenté de r .

Heuristiquement qu'est-ce que cela donne ? Dans le modèle JC69, on place une horloge de taux 1 au dessus de chaque site et quand l'une d'entre elles sonne, l'état du site pour lequel l'horloge a sonné change de manière équiprobable en l'un des

trois états différents de l'état initial. Ensuite toutes les horloges sont réinitialisées et on recommence. Dans le modèle $1+r$, toutes les horloges précédentes sont conservées mais on rajoute des horloges de taux r au dessus de tous les dinucléotides CG . Quand une horloge sonne, on regarde sa nature, si c'est une horloge de taux 1, on procède comme précédemment, si par contre c'est une horloge de taux r le dinucléotide CG saute avec équiprobabilité vers CA ou TG .

Cette simple modification a des conséquences sur le comportement de la chaîne, car ce qui était au départ une séquence avec des sites évoluant indépendamment les uns des autres suivant un noyau markovien plus ou moins sophistiqué, devient une chaîne avec influence du plus proche voisin. A-t-on par exemple une fréquence stationnaire pour ce modèle? Ou bien y a-t-il indépendance des sites quand ils sont assez éloignés dans la séquence? Tout ceci est étudié dans l'article de Gouéré, Bérard et Piau (voir [2]), et il s'avère qu'il y a ergodicité et indépendance des sites s'ils sont séparés par au moins deux nucléotides. Le taux moyen de substitution par unité de temps pour ce modèle est $3 + 2rF(CG)(0)$, où $F(CG)(0)$ est la fréquence stationnaire des dinucléotides CG .

2.2. Fréquences stationnaires pour le modèle $1+r$. — On sait qu'il y a ergodicité pour ce modèle, nous allons établir quelles sont les fréquences stationnaires, ce qui va nous permettre de nous familiariser avec le modèle et de manipuler les générateurs infinitésimaux.

Proposition 2.1. — *Le vecteur*

$$Y(t) = \begin{pmatrix} F(A)(t) \\ F(T)(t) \\ F(C)(t) \\ F(G)(t) \\ F(CG)(t) \end{pmatrix}$$

est solution du système différentiel

$$(2.1) \quad Y'(t) = MY(t) + E$$

avec

$$M = \begin{pmatrix} -4 & 0 & 0 & 0 & r \\ 0 & -4 & 0 & 0 & r \\ 0 & 0 & -4 & 0 & -r \\ 0 & 0 & 0 & -4 & -r \\ 0 & 0 & 1 & 1 & -(8+2r) \end{pmatrix} \quad \text{et} \quad E = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

Démonstration. — Regardons comment évolue la fréquence des sites occupés par des A en fonction du temps t . Comme nous manipulons des noyaux markoviens, nous allons nous placer au temps $t + s$, avec $s > 0$ petit et conditionner par l'état des sites au temps t .

- Si le site était occupé au temps t par un A , il reste un A avec probabilité $F(A|A)(s) = 1 - 3s + o(s)$.
- S'il était occupé par un C ou un T il devient un A avec probabilité $s + o(s)$.
- S'il était occupé par un G , on a besoin de connaître la nature du site situé précédemment,
 - s'il est occupé par un C , CG devient CA avec probabilité $(1+r)s + o(s)$,
 - s'il n'est pas occupé par un C , G devient un A avec probabilité $s + o(s)$.

Nous n'avons regardé que les cas où au plus un nucléotide change d'état, car la probabilité que deux mutations se produisent pendant le temps s est un $o(s)$. On a alors en utilisant la formule de Bayes :

$$\begin{aligned} F(A)(t+s) &= F(A)(t)(1-3s) + F(C)(t)s + F(T)(t)s + F(\bar{C}G)(t)s \\ &\quad + F(CG)(t)(1+r)s + o(s) \end{aligned}$$

d'où, en remarquant que

$$F(\bar{C}G)(t) = F(G)(t) - F(CG)(t) \quad \text{et} \quad F(A) + F(T) + F(C) + F(G) = 1,$$

on obtient

$$F(A)(t+s) = F(A)(t) + (-4F(A)(t) + rF(CG)(t) + 1)s + o(s).$$

En passant $F(A)(t)$ de l'autre côté, en divisant par s et en faisant tendre s vers 0, on obtient l'équation différentielle suivante :

$$(2.2) \quad F(A)'(t) = -4F(A)(t) + rF(CG)(t) + 1$$

Par symétrie du modèle, il vient aisément trois autres équations différentielles qui, rajoutées à (2.2), donnent :

$$\begin{aligned} F(A)'(t) &= -4F(A)(t) + rF(CG)(t) + 1 \\ F(T)'(t) &= -4F(T)(t) + rF(CG)(t) + 1 \\ F(C)'(t) &= -4F(C)(t) - rF(CG)(t) + 1 \\ F(G)'(t) &= -4F(G)(t) - rF(CG)(t) + 1. \end{aligned}$$

Il nous reste à étudier $F(CG)(t)$ pour boucler le système différentiel. Expliquons dans le détail comment on calcule $F(CG|CG)(s)$. Nous avons

$$\begin{aligned} F(CG|CG)(s) &= 1 - F(C\bar{G}|CG)(s) - F(\bar{C}G|CG)(s) + o(s) \\ &= 1 - F(CA|CG)(s) - F(CC|CG)(s) - F(CT|CG)(s) \\ &\quad - F(TG|CG)(s) - F(AG|CG)(s) - F(GG|CG)(s) + o(s) \\ &= 1 - (1+r)s - 2s - (1+r)s - 2s + o(s) \\ &= 1 - 6s - 2rs + o(s) \end{aligned}$$

il vient alors

$$F(CG)(t+s) = F(CG)(t)(1 - 2rs - 6s) + F(C\bar{G})(t)s + F(\bar{C}G)(t)s + o(s)$$

ainsi

$$F(CG)'(t) = F(C)(t) + F(G)(t) - (8 + 2r)F(CG)(t)$$

□

Corollaire 2.2. — *Les fréquences stationnaires pour le modèle $1+r$ sont données par*

$$\begin{aligned} F(A)(0) = F(T)(0) &= \frac{8 + 3r}{32 + 10r}, \\ F(C)(0) = F(G)(0) &= \frac{8 + 2r}{32 + 10r}. \end{aligned}$$

Démonstration. — A stationnarité, les dérivées sont nulles, les fréquences stationnaires sont donc solution de $Mx + E = 0$. La résolution de ce système donne le résultat. □

Remarque. — On a en bonus la fréquence stationnaire des dinucléotides CG , donné par

$$F(CG)(0) = \frac{2}{32 + 10r}$$

2.3. Modèle R/Y + r. — En cours d'étude.

3. Évolutions des modèles

Nous allons donner des estimateurs du temps écoulé lors de l'évolution d'une séquence d'ADN selon nos modèles. Nous supposons qu'un temps relativement long s'est déjà écoulé pour aboutir à la séquence initiale que l'on observe, ce qui nous permet de supposer que les nucléotides sont à fréquence stationnaire. Sur les deux séquences, que peut-on mesurer qui nous fournisse un estimateur du temps écoulé ? On va faire une comparaison site par site et établir des systèmes différentiels nous fournissant divers estimateurs.

3.1. Modèle de Jukes et Cantor. —

Théorème 3.1. — *Dans le modèle JC69, pour tout $N \in \mathcal{A}$, $F(N, N)$ vérifie l'équation différentielle suivante*

$$(3.1) \quad F(N, N)(t)' = -4F(N, N)(t) + F(N)(0).$$

En particulier, si l'on initialise la séquence à l'état où les nucléotides sont à fréquences stationnaires, on a alors une évolution en e^{-4t} des quantités $F(N, N)$ vers $F(N, N)(\infty)$ pour tout $N \in \mathcal{A}$.

Démonstration. — Ici, on n'a pas d'influence du voisinage. Essayons d'établir une équation différentielle vérifiée par $F(C, C)(t)$. On va se placer au temps $t + s$, avec $s > 0$ petit, et tâcher d'évaluer le comportement de $F(C, C)(t + s) - F(C, C)(t)$ quand s tend vers 0. Pour cela on va utiliser la propriété de Markov, qui nous dit que ce que l'on observe en un site au temps $t + s$ peut être conditionné par l'état du site au temps t :

$$\begin{aligned} F(C, C)(t + s) &= F(C, C, C)(t, s) + F(C, \bar{C}, C)(t, s) \\ &= F(C, C)(t)F(C|C)(s) + F(C, \bar{C})(t)F(C|\bar{C})(s) \\ &= F(C, C)(t)(1 - 3s) + F(C, \bar{C})s + o(s) \\ &= F(C, C)(t)(1 - 3s) + (F(C)(0) - F(C, C)(t))s + o(s) \\ &= F(C, C)(t) - 4F(C, C)(t)s + F(C)(0)s + o(s) \end{aligned}$$

ainsi

$$F(C, C)(t + s) = F(C, C)(t) + (-4F(C, C)(t) + F(C)(0))s + o(s).$$

On a donc établi l'équation différentielle suivante :

$$F(C, C)(t)' = -4F(C, C)(t) + F(C)(0)$$

On peut remarquer que cette équation différentielle est vérifiée par $F(A, A)$, $F(T, T)$ et $F(G, G)$, ceci ne sera pas le cas dans les modèles avec influence du voisinage.

La résolution de (3.1) nous donne :

$$F(N, N)(t) = \frac{F(N)(0)}{4} (1 + 3e^{-4t}) = \frac{1}{16} (1 + 3e^{-4t}).$$

□

Si l'on appelle L la longueur de la chaîne et $(X_i(t))_{i=1}^L$ la nature des sites de la chaîne au temps t , on tire de la relation précédente l'estimateur suivant :

$$\hat{t} = \frac{1}{4} \ln \left(\frac{3}{16\hat{F}(C, C)(t) - 1} \right) \quad \text{avec} \quad \hat{F}(C, C)(t) = \frac{1}{L} \sum_{k=1}^L \mathbf{1}_{[X_k(0)=X_k(t)=C]}$$

Notre estimateur n'est valable que si $\hat{F}(C, C)(t)$ est strictement supérieur à $F(C, C)(\infty) = \frac{1}{16}$ et il explose quand $\hat{F}(C, C)(t)$ est proche de $\frac{1}{16}$, ceci est normal car si la fréquence est proche de $\frac{1}{16}$ qui est la valeur de $F(C, C)(\infty)$, il est logique de penser que le temps écoulé est long.

3.2. Modèle $1+r$. — On va utiliser la même technique, mais cette fois plusieurs quantités interviennent pour établir un système différentiel. Ceci est dû au fait que l'on doit regarder la nature du nucléotide situé à la droite d'un site occupé par un C ou à situé à la gauche d'un site occupé par un G.

3.2.1. Obtention de systèmes différentiels. — Énonçons le théorème suivant :

Théorème 3.2. — *Dans le modèle $1+r$, le vecteur*

$$U(t) = \begin{pmatrix} F(C, C)(t) \\ F(C^*, CG)(t) \\ F(C^*, \bar{C}G)(t) \end{pmatrix}$$

est solution du système différentiel

$$(3.2) \quad U'(t) = M_1 U(t) + F(C)(0)B_1$$

avec

$$M_1 = \begin{pmatrix} -4 & -r & 0 \\ 1 & -(7+2r) & 1 \\ -1 & 3+r & -5 \end{pmatrix} \quad \text{et} \quad B_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

tandis que le vecteur

$$V(t) = \begin{pmatrix} F(A, A)(t) \\ F(A, G)(t) \\ F(*A, CG)(t) \\ F(*A, \bar{C}G)(t) \end{pmatrix}$$

est solution du système différentiel

$$(3.3) \quad V'(t) = M_2 V(t) + F(A)(0)B_2$$

avec

$$M_2 = \begin{pmatrix} -4 & 0 & r & 0 \\ 0 & -4 & -r & 0 \\ 0 & 1 & -(7+2r) & 1 \\ 0 & -1 & 3+r & -5 \end{pmatrix} \quad \text{et} \quad B_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

Ce théorème a été établi en étudiant les variations de diverses quantités. Nous avons commencé par $F(C, C)$ et constaté que son évolution était couplée avec d'autres quantités, que nous avons étudiées afin de fermer le système. Nous avons obtenus le lemme suivant :

Lemme 3.3. — *Les quantités $F(C, C)$, $F(C^*, CG)$ et $F(C^*, \bar{C}G)$ vérifient les équations différentielles suivantes :*

$$(3.4) \quad F(C, C)'(t) = -4F(C, C)(t) - rF(C^*, CG)(t) + F(C)(0)$$

$$(3.5) \quad F(C^*, CG)'(t) = F(C, C)(t) - (7+2r)F(C^*, CG)(t) + F(C^*, \bar{C}G)(t)$$

$$(3.6) \quad F(C^*, \bar{C}G)'(t) = -F(C, C)(t) + (r+3)F(C^*, CG)(t) - 5F(C^*, \bar{C}G)(t) + F(C)(0)$$

Démonstration. — Regardons l'évolution de $F(C, C)(t)$. Soit $s > 0$ petit, on a alors :

$$\begin{aligned} F(C, C)(t+s) &= F(C, C, C)(t, s) + F(C, \bar{C}, C)(t, s) \\ &= F(C^*, CG, C^*)(t, s) + F(C^*, C\bar{G}, C^*)(t, s) + F(C, \bar{C})(t)s + o(s) \\ &= F(C^*, CG, CG)(t, s) + F(C^*, CG, C\bar{G})(t, s) \\ &\quad + F(C^*, C\bar{G})(t)(1-3s) + F(C, \bar{C})(t)s + o(s) \end{aligned}$$

On a déjà vu que

$$F(CG|CG)(s) = 1 - F(C\bar{G}|CG)(s) - F(\bar{C}G|CG)(s) + o(s) = 1 - 6s - 2rs + o(s)$$

il vient alors

$$\begin{aligned} F(C, C)(t+s) &= F(C^*, CG)(t)(1-6s-2rs) + F(C^*, CG)(t)(rs+3s) \\ &\quad + F(C^*, C\bar{G})(t)(1-3s) + F(C, \bar{C})(t)s + o(s) \\ &= F(C^*, CG)(t) + F(C^*, C\bar{G})(t) - (3+r)F(C^*, CG)(t)s \\ &\quad - 3F(C^*, C\bar{G})(t)s + (F(C)(0) - F(C, C)(t))s + o(s) \\ &= F(C, C)(t) - 4F(C, C)(t)s - rF(C^*, CG)(t)s + F(C)(0)s + o(s). \end{aligned}$$

En passant $F(C, C)(t)$ de l'autre côté, en divisant par s et en faisant tendre s vers 0, on obtient l'équation différentielle (3.4).

Pour la quantité $F(C^*, CG)$, on a cette fois affaire à des dinucléotides. Comme au cours du temps s , la probabilité que plus de deux mutations aient lieu est un $o(s)$, on ne va conditionner au temps t que par des dinucléotides qui diffèrent en au plus un site de CG . Il vient avec $s > 0$ petit

$$\begin{aligned} F(C^*, CG)(t+s) &= F(C^*, CG, CG)(t, s) + F(C^*, C\bar{G}, CG)(t, s) \\ &\quad + F(C^*, \bar{C}G, CG)(t, s) + o(s) \\ &= F(C^*, CG)(t)(1-6s-2rs) + F(C^*, C\bar{G})(t)s \\ &\quad + F(C^*, \bar{C}G)(t)s + o(s) \\ &= F(C^*, CG)(t) + F(C, C)(t)s - (7+2r)F(C^*, CG)(t)s \\ &\quad + F(C^*, \bar{C}G)(t)s + o(s). \end{aligned}$$

En passant $F(C^*, CG)(t)$ de l'autre côté, en divisant par s et en faisant tendre s vers 0, on obtient l'équation différentielle (3.5).

L'équation différentielle (3.6) s'obtient avec un raisonnement similaire. \square

Les équations (3.4), (3.5) et (3.6) nous permettent d'établir le système (3.2).

Nous avons procédé avec la quantité $F(A, A)$ comme avec $F(C, C)$. Nous allons vérifier que cette quantité ne fait pas intervenir le même système différentiel, alors que dans le modèle de Jukes et Cantor $F(A, A)$ et $F(C, C)$ sont solution de la même équation différentielle.

Lemme 3.4. — *Les quantités $F(A, A)$, $F(*A, CG)$, $F(*A, C\bar{G})$ et $F(A, G)$ vérifient les équations différentielles suivantes :*

$$(3.7) \quad F(A, A)'(t) = -4F(A, A)(t) + rF(*A, CG)(t) + F(A)(0)$$

$$(3.8) \quad F(*A, CG)'(t) = -(7+2r)F(*A, CG)(t) + F(*A, C\bar{G})(t) + F(A, G)(t)$$

$$(3.9) \quad F(*A, C\bar{G})'(t) = (3+r)F(*A, CG)(t) - 5F(*A, C\bar{G})(t) - F(A, G)(t) + F(A)(0)$$

$$(3.10) \quad F(A, G)'(t) = -rF(*A, CG)(t) - 4F(A, G)(t) + F(A)(0).$$

Démonstration. — On applique le même raisonnement que pour le lemme 3.3. \square

Les équations (3.7), (3.8), (3.9) et (3.10) nous permettent d'établir le système (3.3).

Remarque. — Le vecteur

$$\begin{pmatrix} F(A, G)(t) \\ F(*A, CG)(t) \\ F(*A, C\bar{G})(t) \end{pmatrix}$$

est solution du système différentiel

$$(3.11) \quad \tilde{U}' = M_1 \tilde{U} + F(A)(0)B_1.$$

3.2.2. Exploitation des systèmes différentiels obtenus. — Nous allons voir que le modèle $1+r$ diffère du modèle de Jukes et Cantor par l'évolution des quantités $F(*N, CG)$, avec $N \in Y$, ou $F(N*, CG)$ avec $N \in R$. Nous allons par exemple démontrer que $F(C*, CG)$ converge plus vite vers $F(C*, CG)(\infty)$ dans le modèle $1+r$ que dans le modèle JC69, mais pas $F(C, C)$.

Théorème 3.5. — Notons $\omega = \sqrt{r^2 + 2r + 4}$. Dans le modèle $1+r$, partant de l'état où les nucléotides sont à fréquences stationnaires, la quantité $F(C*, CG)$ converge vers $F(C*, CG)(\infty)$ en $e^{(-6-r+\omega)t}$, tandis que la quantité $F(C, C)$ converge vers $F(C, C)(\infty)$ en e^{-4t} .

Avant de démontrer le théorème 3.5, il vient le

Corollaire 3.6. — A taux moyen de substitution par unité de temps égal, la quantité $F(C*, CG)$ converge plus vite vers $F(C*, CG)(\infty)$ dans le modèle $1+r$ que dans le modèle JC69.

Démonstration. — Tout d'abord, il nous faut ajuster λ dans le modèle JC69 pour avoir le même taux moyen de substitution par unité de temps que dans le modèle $1+r$. Il vient

$$\lambda_r = 1 + \frac{2r}{3(16+5r)} = \frac{48+17r}{3(16+5r)}.$$

Dans le modèle JC69, la convergence vers l'équilibre est en $e^{-4\lambda_r t}$. En effet par indépendance des sites, il vient

$$F(C*, CG)(t) = F(C, C)(t)F(*, G)(t) = \frac{1}{4} \cdot \frac{1}{16} (1 + 3e^{-4\lambda_r t}).$$

Dans le modèle $1+r$, la convergence est en $e^{-(6+r-\omega)t}$ d'après le théorème 3.5. Or après calculs, on a

$$4\lambda_r \leq 6+r-\omega \iff r(210r^2 + 1696r + 3072) \geq 0.$$

\square

Remarque. — La quantité $6+r-\omega$ vue comme une fonction de r est croissante, elle vaut 4 en 0 et elle tend vers 6 en $+\infty$.

Démonstration du théorème 3.5. — Commençons par étudier la matrice M_1 . Son polynôme caractéristique est

$$(X+4)(X+6+r+\omega)(X+6+r-\omega) \text{ avec } \omega = \sqrt{r^2 + 2r + 4}$$

On constate que -4 est la plus grande valeur propre de M_1 . Les vecteurs

$$u = \begin{pmatrix} r \\ 2+r-\omega \\ \omega-2 \end{pmatrix}, \quad v = \begin{pmatrix} r \\ 2+r+\omega \\ -\omega-2 \end{pmatrix} \quad \text{et} \quad w = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

sont vecteurs propres de M_1 associés aux valeurs propres respectives $-6 - r + \omega$, $-6 - r - \omega$ et -4 .

Comme toutes les valeurs propres de M_1 sont négatives, U_{sol} une solution du système différentiel (3.2), vérifie $U'_{sol}(\infty) = 0$, d'où $F(C)(0)B_1 = -M_1 U_{sol}(\infty)$. Le système (3.2) peut donc se réécrire $(U - U(\infty))' = M_1(U - U(\infty))$. La solution de ce système vérifiant $U_{sol}(0) = U(0)$ est $U(t) = e^{M_1 t}(U(0) - U(\infty)) + U(\infty)$. Mais nous nous sommes placés dans le cas où les fréquences sont stationnaires, on a donc

$$U(0) = \begin{pmatrix} F(C, C)(0) \\ F(C^*, CG)(0) \\ F(C^*, \bar{C}G)(0) \end{pmatrix} = \begin{pmatrix} F(C)(0) \\ F(CG)(0) \\ 0 \end{pmatrix}$$

$$U(\infty) = \begin{pmatrix} F(C)(0)^2 \\ F(C)(0)F(CG)(0) \\ F(C)(0)(F(C)(0) - F(CG)(0)) \end{pmatrix}$$

ainsi,

$$U(0) - U(\infty) = \frac{3+r}{(16+5r)^2} \begin{pmatrix} 4(4+r) \\ 4 \\ -(4+r) \end{pmatrix}$$

La décomposition dans la base (u, v, w) de $U(0) - U(\infty)$ donne

$$U(0) - U(\infty) = \alpha u + \beta v + \gamma w$$

avec

$$\alpha = \frac{3+r}{4r\omega^2(16+5r)^2} (64 + 44r + 22r^2 + 3r^3 + 32\omega + 14r\omega + 3r^2\omega)$$

$$\beta = \frac{3+r}{4r\omega^2(16+5r)^2} (64 + 44r + 22r^2 + 3r^3 - 32\omega - 14r\omega - 3r^2\omega)$$

$$\gamma = \frac{3+r}{2(16+5r)}$$

On a ainsi,

$$e^{M_1 t}(U(0) - U(\infty)) = \alpha e^{(-6-r+\omega)t}u + \beta e^{(-6-r-\omega)t}v + \gamma e^{-4t}w$$

on en déduit

$$F(C^*, CG)(t) = \alpha(2+r-\omega)e^{(-6-r+\omega)t} + \beta(2+r+\omega)e^{(-6-r-\omega)t} + \frac{4+r}{(16+5r)^2}$$

et

$$F(C, C)(t) = \alpha r e^{(-6-r+\omega)t} + \beta r e^{(-6-r-\omega)t} + \gamma e^{-4t} + \left(\frac{4+r}{16+5r}\right)^2.$$

On peut vérifier heuristiquement que ce que l'on trouve est cohérent avec le modèle JC69. En effet, quand r est proche de 0, le modèle $1+r$ se comporte sensiblement comme le modèle JC69. Or, nous avons

$$(2+r-\omega)\alpha = \frac{3}{64} + o(1)$$

$$(2+r+\omega)\beta = o(r)$$

$$\frac{4+r}{(16+5r)^2} = \frac{1}{64} + o(1).$$

On a donc quand r est proche de 0

$$F(C^*, CG)(t) \approx \frac{3}{64}e^{-4t} + \frac{1}{64}.$$

Or dans le modèle JC69, on a

$$F(C^*, CG)(t) = \frac{1}{4} \cdot \frac{1}{16} (1 + 3e^{-4t}),$$

ce qui est bien cohérent avec notre résultat. De même, nous avons

$$\begin{aligned} r\alpha &= \frac{3}{32} + o(1) \\ r\beta &= o(r^2) \\ \gamma &= \frac{3}{32} + o(1) \\ \left(\frac{4+r}{16+5r}\right)^2 &= \frac{1}{64} + o(1). \end{aligned}$$

On a donc quand r est proche de 0

$$F(C, C)(t) \approx \frac{3}{16}e^{-4t} + \frac{1}{16}.$$

Or dans le modèle JC69, on a

$$F(C, C)(t) = \frac{1}{16}(1 + 3e^{-4t})$$

ce qui est bien cohérent avec notre résultat. \square

Théorème 3.7. — *Dans le modèle $1+r$, partant de l'état où les nucléotides sont à fréquences stationnaires, la quantité $F(*A, CG)$ converge vers $F(*A, CG)(\infty)$ en $e^{(-6-r+\omega)t}$, tandis que la quantité $F(A, A)$ converge vers $F(A, A)(\infty)$ en e^{-4t} .*

Démonstration. — Utilisons le fait que nous avons déjà travaillé sur la matrice M_1 . \tilde{U}_{sol} une solution du système différentiel (3.11), vérifie $\tilde{U}'_{sol}(\infty) = 0$, d'où $F(A)(0)B_1 = -M_1\tilde{U}_{sol}(\infty)$. Le système (3.11) peut donc se réécrire $(\tilde{U} - \tilde{U}(\infty))' = M_1(\tilde{U} - \tilde{U}(\infty))$. La solution de ce système vérifiant $\tilde{U}_{sol}(0) = \tilde{U}(0)$ est $\tilde{U}(t) = e^{M_1 t}(\tilde{U}(0) - \tilde{U}(\infty)) + \tilde{U}(\infty)$. On a

$$\begin{aligned} \tilde{U}(0) &= \begin{pmatrix} F(A, G)(0) \\ F(*A, CG)(0) \\ F(*A, C\bar{G})(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ F(CA)(0) \end{pmatrix} \\ \tilde{U}(\infty) &= \begin{pmatrix} F(A)(0)F(G)(0) \\ F(A)(0)F(CG)(0) \\ F(A)(0)(F(C)(0) - F(CG)(0)) \end{pmatrix}. \end{aligned}$$

Le terme $F(CA)(0)$ est inconnu pour l'instant, mais nous pouvons lever rapidement cette indétermination. En effet, on a pour $s > 0$ petit

$$\begin{aligned} F(CA)(t+s) &= F(CA)(t)F(CA|CA)(s) + F(C\bar{A})(t)F(CA|C\bar{A})(s) \\ &\quad + F(\bar{C}A)(t)F(CA|\bar{C}A)(s) + o(s) \\ &= F(CA)(t)(1-6s) + F(C\bar{A})(t)s + F(CG)(t)rs \\ &\quad + F(\bar{C}A)(t)s + o(s) \\ &= F(CA)(t) - 8F(CA)(t)s + rF(CG)(t)s + F(C)(t)s \\ &\quad + F(A)(t)s + o(s) \end{aligned}$$

ainsi

$$F(CA)'(t) = F(A)(t) + F(C)(t) - 8F(CA)(t) + rF(CG)(t)$$

A stationnarité, la dérivé est nulle et les autres fréquences stationnaires sont connues, il sort donc

$$F(CA)(0) = \frac{16+7r}{8(32+10r)}.$$

Ainsi on a

$$\tilde{U}(0) - \tilde{U}(\infty) = \frac{1}{4(32+10r)^2} \begin{pmatrix} -8(8+3r)(4+r) \\ -8(8+3r) \\ 64+56r+11r^2 \end{pmatrix}$$

Décomposons $\tilde{U}(0) - \tilde{U}(\infty)$ dans la base (u, v, w) . On a

$$\tilde{U}(0) - \tilde{U}(\infty) = \tilde{\alpha}u + \tilde{\beta}v + \tilde{\gamma}w$$

avec

$$\begin{aligned}\tilde{\alpha} &= -\frac{1}{64r\omega^2(16+5r)^2} (1024 + 1024r + 564r^2 + 154r^3 + 13r^4 \\ &\quad + 512\omega + 384r\omega + 106\omega r^2 + 13\omega r^3) \\ \tilde{\beta} &= -\frac{1}{64r\omega^2(16+5r)^2} (1024 + 1024r + 564r^2 + 154r^3 + 13r^4 \\ &\quad - 512\omega - 384r\omega - 106\omega r^2 - 13\omega r^3) \\ \tilde{\gamma} &= -\frac{16+7r}{32(16+5r)}.\end{aligned}$$

d'où

$$e^{M_1 t}(\tilde{U}(0) - \tilde{U}(\infty)) = \tilde{\alpha}e^{(-6-r+\omega)t}u + \tilde{\beta}e^{(-6-r-\omega)t}v + \tilde{\gamma}e^{-4t}w.$$

On en déduit

$$F(*A, CG)(t) = \tilde{\alpha}e^{(-6-r+\omega)t}(2+r-\omega) + \tilde{\beta}e^{(-6-r-\omega)t}(2+r+\omega) + \frac{2(8+3r)}{(32+10r)^2}.$$

Avant de poursuivre, on peut vérifier de la même manière que pour $F(C*, CG)$, que ce que l'on trouve est cohérent avec le modèle JC69. Quand r est proche de 0, nous avons

$$\begin{aligned}(2+r-\omega)\tilde{\alpha} &= -\frac{1}{64} + o(1) \\ (2+r+\omega)\tilde{\beta} &= o(r) \\ \frac{2(8+3r)}{(32+10r)^2} &= \frac{1}{64} + o(1).\end{aligned}$$

On a donc quand r est proche de 0

$$F(*A, CG)(t) \approx -\frac{1}{64}e^{-4t} + \frac{1}{64}.$$

Or dans le modèle JC69, on a

$$F(*A, CG)(t) = F(*, C)(t)F(A, G)(t) = \frac{1}{4} \cdot \frac{1}{16} (1 - e^{-4t}),$$

ce qui est bien cohérent avec notre résultat. On peut donc à présent résoudre l'équation différentielle (3.7). La fonction

$$t \mapsto e^{-4t}$$

est solution de l'équation homogène et par la méthode de variation de la constante, on obtient comme solution particulière la fonction

$$t \mapsto -r\tilde{\alpha}e^{(-6-r+\omega)t} - r\tilde{\beta}e^{(-6-r-\omega)t} + \frac{r(8+3r)}{2(32+10r)^2} + \frac{8+3r}{4(32+10r)}.$$

Or, on a

$$\frac{r(8+3r)}{2(32+10r)^2} + \frac{8+3r}{4(32+10r)} = \frac{(8+3r)^2}{(32+10r)^2},$$

il existe donc une constante $C \in \mathbf{R}$ telle que

$$F(A, A)(t) = Ce^{-4t} - r\tilde{\alpha}e^{(-6-r+\omega)t} - r\tilde{\beta}e^{(-6-r-\omega)t} + \frac{(8+3r)^2}{(32+10r)^2}.$$

Or en $t = 0$, il vient

$$\frac{8+3r}{32+10r} = C - r\tilde{\alpha} - r\tilde{\beta} + \frac{(8+3r)^2}{(32+10r)^2}$$

Ce qui nous donne

$$C = \frac{31r + 80}{32(16 + 5r)} \neq 0.$$

Comme précédemment, vérifions que cette expression est cohérente. Quand r est proche de 0, on a

$$\begin{aligned} C &= \frac{5}{32} + o(1) \\ -r\tilde{\alpha} &= \frac{1}{32} + o(1) \\ -r\tilde{\beta} &= o(r^2) \\ \frac{(8 + 3r)^2}{(32 + 10r)^2} &= \frac{1}{16} + o(1). \end{aligned}$$

On a donc quand r est proche de 0

$$F(A, A)(t) \approx \frac{3}{16}e^{-4t} + \frac{1}{16}.$$

Or dans le modèle JC69, on a

$$F(A, A)(t) = \frac{1}{16}(1 + 3e^{-4t}),$$

ce qui montre que notre expression est cohérente. □

Nous avons obtenu des expressions pour diverses quantités en fonction du temps dans le modèle $1 + r$. Notre but est d'en tirer un estimateur du temps écoulé entre une séquence actuelle et sa séquence ancestrale. À partir de ceci, nous allons essayer de donner une distance entre deux séquences actuelles d'ADN issues d'une même séquence ancestrale, distance qui nous servira pour construire une phylogénie pour une collection de séquences actuelles.

3.3. Modèle $\mathbf{R}/\mathbf{Y} + \mathbf{r}$. — En cours d'étude

Proposition 3.8. — *La quantité $F(C, C)$ vérifie l'équation différentielle suivante :*

$$(3.12) \quad \begin{aligned} F(C, C)'(t) &= v_C F(C, A)(t) + w_C F(C, T)(t) \\ &\quad - (v_A + w_T + v_G) F(C, C)(t) \\ &\quad + v_C F(C, G)(t) - r F(C^*, CG)(t). \end{aligned}$$

Démonstration. — Soit $s > 0$ petit, il vient

$$\begin{aligned} F(C, C)(t + s) &= F(C, A, C)(t, s) + F(C, T, C)(t, s) + F(C, C, C)(t, s) \\ &\quad + F(C, G, C)(t, s) \\ &= F(C, A)(t)v_C s + F(C, T)(t)w_C s + F(C, G)(t)v_C s \\ &\quad + F(C^*, CG, C^*)(t, s) + F(C^*, C\bar{G}, C^*)(t, s) + o(s) \\ &= F(C^*, CG)(t)F(CG|CG)(s) + F(C^*, CG)(t)F(C\bar{G}|CG)(t, s) \\ &\quad + F(C^*, CA, C^*)(t, s) + F(C^*, CT, C^*)(t, s) + F(C^*, CC, C^*)(s) \\ &\quad + F(C, A)(t)v_C s + F(C, T)(t)w_C s + F(C, G)(t)v_C s + o(s) \\ &= F(C^*, CG)(t)(F(CG|CG)(s) + F(C\bar{G}|CG)(s) \\ &\quad + F(C^*, CA)(t)F(C^*|CA)(s) + F(C^*, CT)(t)F(C^*|CT)(s) \\ &\quad + F(C^*, CC)(t)F(C^*|CC)(s) \\ &\quad + F(C, A)(t)v_C s + F(C, T)(t)w_C s + F(C, G)(t)v_C s + o(s) \end{aligned}$$

mais

$$\begin{aligned} F(CG|CG)(s) + F(C\bar{G}|CG)(s) &= 1 - F(\bar{C}G|CG)(s) + o(s) \\ &= 1 - v_A s - (w_T + r)s - v_G s + o(s) \end{aligned}$$

et

$$F(C * |CA)(s) = F(C * |CT)(s) = F(C * |CC)(s) = 1 - s(v_A + w_T + v_G)$$

d'où

$$\begin{aligned} F(C, C)(t + s) &= F(C*, CG)(t)(1 - v_A s - (w_T + r)s - v_G s) \\ &\quad + (F(C, C)(t) - F(C*, CG)(t))(1 - s(v_A + w_T + v_G)) \\ &\quad + F(C, A)(t)v_C s + F(C, T)(t)w_C s + F(C, G)(t)v_C s + o(s) \\ &= F(C, C)(t)(1 - s(v_A + w_T + v_G)) \\ &\quad + F(C, A)(t)v_C s + F(C, T)(t)w_C s + F(C, G)(t)v_C s \\ &\quad - F(C*, CG)(t)rs + o(s). \end{aligned}$$

En passant $F(C, C)(t)$ de l'autre côté, en divisant par s et en faisant tendre s vers 0, on obtient l'équation différentielle (3.12). \square

Le rôle de G étant similaire à celui de C , on obtient

Proposition 3.9. — La quantité $F(C, G)$ vérifie l'équation différentielle suivante :

$$(3.13) \quad F(C, G)'(t) = w_G F(C, A)(t) + v_G F(C, T)(t) + v_G F(C, C)(t) - (w_A + v_T + v_C)F(C, G)(t) - rF(*C, CG)(t).$$

Qu'en est-il de la quantité $F(C, A)$?

Proposition 3.10. — La quantité $F(C, A)$ vérifie l'équation différentielle :

$$(3.14) \quad F(C, A)'(t) = -(v_T + v_C + w_G)F(C, A)(t) + v_A F(C, T)(t) + v_A F(C, C)(t) + w_A F(C, G)(t) + rF(*C, CG)(t).$$

Démonstration. —

$$\begin{aligned} F(C, A)(t + s) &= F(C, A, A)(t, s) + F(C, T, A)(t, s) + F(C, C, A)(t, s) \\ &\quad + F(C, G, A)(t, s) \\ &= F(C, A)(t)(1 - s(v_T + v_C + w_G)) + F(C, T)(t)v_A s \\ &\quad + F(C, C)(t)v_A s + F(*C, CG, *A) + F(*C, \bar{C}G, *A) + o(s) \end{aligned}$$

or

$$\begin{aligned} F(*C, CG, *A) &= F(*C, CG, CA) + F(*C, CG, \bar{C}A) \\ &\quad + F(*C, CG)(w_A + r)s + o(s) \end{aligned}$$

et

$$\begin{aligned} F(*C, \bar{C}G, *A) &= F(*C, AG, *A) + F(*C, TG, *A) + F(*C, GG, *A) \\ &= F(*C, AG, AA) + F(*C, TG, TA) + F(*C, GG, GA) + o(s) \\ &= F(*C, AG)w_A s + F(*C, TG)w_A s + F(*C, GG)w_A s + o(s) \end{aligned}$$

d'où

$$\begin{aligned} F(C, A)(t + s) &= F(C, A)(t)(1 - s(v_T + v_C + w_G)) + F(C, T)(t)v_A s \\ &\quad + F(C, C)(t)v_A s + F(C, G)w_A s + F(*C, CG)(t)rs + o(s). \end{aligned}$$

En passant $F(C, A)(t)$ de l'autre côté, en divisant par s et en faisant tendre s vers 0, on obtient l'équation différentielle (3.14). \square

Passons à $F(C, T)$, par symétrie du rôle de A et T , on a immédiatement

Proposition 3.11. — *La quantité $F(C, T)$ vérifie l'équation différentielle :*

$$(3.15) \quad F(C, T)'(t) = v_T F(C, A)(t) - (v_A + w_C + v_G)F(C, T)(t) \\ + w_T F(C, C)(t) + v_T F(C, G)(t) + rF(C^*, CG)(t).$$

Voyons à présent la quantité $F(C^*, CG)$

$$\begin{aligned} F(C^*, CG)(t + s) &= F(C^*, CG, CG)(t, s) + F(C^*, C\bar{G}, CG)(t, s) \\ &\quad + F(C^*, \bar{C}G, CG)(t, s) + o(s) \\ &= F(C^*, CG)(t)(1 - (w_A + r)s - v_T s - v_C s \\ &\quad - v_A s - (w_T + r)s - v_G s) + F(C^*, CA)(t)w_G s \\ &\quad + F(C^*, CT)(t)v_G s + F(C^*, CC)(t)v_G s + F(C^*, AG)(t)v_C s \\ &\quad + F(C^*, TG)(t)w_T s + F(C^*, GG)(t)v_T s + o(s) \end{aligned}$$

PARTIE II

ARBRE PHYLOGÉNÉTIQUE OU DENDROGRAMME

Cette partie est largement inspirée du livre de Barthélémy et Guénoche (voir [1]).

Nous avons tous en tête l'image d'un arbre généalogique ou suivant notre culture en biologie, un vague souvenir d'arbre représentant les liens entre les reptiles, les poissons, les grands singes et l'homme. Cette partie est faite pour formaliser la notion d'arbre et définir ce qu'est un arbre phylogénétique. Rappelons que notre objectif est de reconstruire l'arbre phylogénétique de séquences d'ADN actuelles issues d'une même séquence ancestrale. L'aspect reconstruction n'est abordée que dans la partie suivante, nous ne posons ici que les bases et les notions qui nous seront utiles pour la suite. Ceci en fait une partie assez formelle mais nécessaire pour la suite.

4. Quelques rappels sur la théorie des graphes

Les graphes permettent de représenter des liaisons entre des objets. Formellement, cela donne

Définition 4.1. — *Un graphe est un couple $G = (S, A)$ formé d'un ensemble fini S et d'un ensemble A de parties à 2 éléments de S . Les éléments de S sont appelés sommets et ceux de A arêtes. L'arête $\{u, v\}$ sera notée uv et on dira que u et v sont adjacents ou qu'ils sont extrémités de l'arête uv .*

Définition 4.2. — *L'ensemble des sommets adjacents à un sommet u est appelé voisinage de u et est noté $V(u)$. Le nombre $\text{deg}(u)$ des éléments de $V(u)$ est appelé le degré du sommet u .*

Définition 4.3. — *Un chemin de G entre les sommets u et v est une suite finie*

$$c : u = u_0 u_1 u_2 \dots u_{p-1} u_p = v$$

telle que pour $0 \leq i < p$, $u_i u_{i+1} \in A$. De plus on suppose que chaque arête $u_i u_{i+1}$ n'intervient qu'une fois dans c . L'entier p est appelé longueur du chemin c , les sommets u et v sont les extrémités du chemin.

Ainsi, une arête est un chemin de longueur 1 et un sommet, un chemin de longueur nulle.

Définition 4.4. — Un cycle est un chemin dont les extrémités coïncident.

Définition 4.5. — Un graphe G est dit connexe si pour chaque paire de sommets de G , il existe un chemin entre ces sommets.

Définition 4.6. — Le sous graphe d'un graphe G induit par un sous ensemble S' de S admet S' pour ensemble de sommets et ses arêtes sont toutes les arêtes de G dont les deux extrémités sont dans S' .

Définition 4.7. — Si A' est un sous ensemble des arêtes du graphe $G = (S, A)$, le graphe $G' = (S, A')$ est appelé le graphe partiel de G induit par A' .

Nous venons de formaliser des liaisons entre des objets, mais nous n'avons pas cherché à privilégier des liaisons par des poids ou des intensités. Or dans les arbres phylogénétiques, la longueur de l'arête entre l'ancêtre et les descendants représente le temps, nous allons donc introduire une nouvelle définition, celle de graphe valué.

Définition 4.8. — Un graphe valué est un couple (G, L) formé d'un graphe $G = (S, A)$ et d'une fonction L , à valeurs réelles strictement positives, définie sur A . $L(uv)$ est appelée la longueur ou la valuation de l'arête uv . Dans un graphe valué (G, L) , la longueur $L(c)$ d'un chemin c est la somme des longueurs des arêtes qui le composent.

5. Arbres et X-arbres

5.1. Arbres. — Donnons la définition d'un arbre et le vocabulaire adapté à ce type de graphe.

Définition 5.1. — Un arbre est un graphe connexe et sans cycle. Dans un arbre, tout sommet de degré 1 est appelé une feuille. Les autres sommets sont appelés noeuds. Nous appellerons arbre binaire un arbre dont tous les noeuds sont de degré 3.

Voici un lemme qui va nous être utile pour donner une meilleure caractérisation des arbres par la suite.

Lemme 5.2. — Un arbre avec au moins deux sommets admet au moins deux feuilles.

Démonstration. — Raisonnons par l'absurde. Soit G un arbre avec au moins deux sommets et admettant une seule feuille u_0 , notons u_1 le sommet adjacent de u_0 . Pour $p \geq 1$ on peut définir u_{p+1} un sommet adjacent de u_p et différent de u_{p-1} . En effet, si $u_p = u_0$, c_p le chemin $u_0u_1 \dots u_p$ est un cycle de G ce qui contredit le fait que G est un arbre, on en déduit donc que u_p n'est pas une feuille et que ce qui précède est possible. Soit c_p la suite des chemins définis comme précédemment, comme G est sans cycle les sommets de c_p sont distincts deux à deux et ce pour tout p . G possède donc une infinité de sommets, ce qui est absurde. Dans le cas où G n'a pas de feuilles, on choisit un sommet quelconque de G et on applique le même raisonnement. \square

Proposition 5.3. — Les assertions suivantes sont équivalentes :

- (i) $G = (S, A)$ est un arbre.
- (ii) G est sans cycle et $|S| = |A| + 1$.
- (iii) G est connexe et $|S| = |A| + 1$.
- (iv) Deux sommets quelconques de G sont toujours reliés par un chemin et un seul.
- (v) G est connexe et, pour toute arête uv de G , le graphe partiel de G induit par $A \setminus \{uv\}$ n'est plus connexe.

Démonstration. — On montre que (i) implique (ii) en effectuant une récurrence sur le nombre n des sommets de G . Le résultat est clair pour $n = 1$. Soit $n \geq 2$, supposons que le résultat soit vrai pour $n - 1$ et considérons un arbre G possédant n sommets. Comme $n \geq 2$, en vertu du lemme 5.2, G admet au moins une feuille u . Désignons par uv , l'unique arête issue de u et considérons le sous graphe H de G induit par $S \setminus \{u\}$. H est connexe et sans cycle. On peut donc lui appliquer l'hypothèse de récurrence, H possède $n - 1$ sommets et $n - 2$ arêtes. Mais on passe de H à G en ajoutant le sommet u et l'arête uv , G possède donc n sommets et $n - 1$ arêtes.

À présent, montrons que (ii) est équivalent à (iii). Supposons G sans cycle et $|S| = |A| + 1$. Soit k le nombre de composantes connexes de G et pour $1 \leq i \leq k$, notons $G_i = (S_i, A_i)$ les sous graphes connexes de G . Comme G est sans cycle, chaque G_i est un arbre, on a donc pour $1 \leq i \leq k$, $|S_i| = |A_i| + 1$. Ces ensembles étant disjoints deux à deux, ceci entraîne $|S| = |A| + k$. On a donc $k = 1$ et G est connexe. Réciproquement, supposons G connexe et $|S| = |A| + 1$. On construit une suite $G = G_0, G_1, \dots, G_p$ de graphes partiels connexes de G de la manière suivante :

- si G_i possède un cycle $u = u_0u_1u_2 \dots u_{l-1}u_l = u$, on construit G_{i+1} en supprimant l'arête $u_{l-1}u_l$. Ainsi, si G_i est connexe, G_{i+1} reste connexe.
- Si G_i ne possède pas de cycle, on arrête la construction.

Comme A est fini, le nombre de fois où la construction est possible est fini, la suite s'arrête donc. Le graphe G_p est un graphe connexe et sans cycle, c'est donc un arbre, d'où $|S| = |S_p| = |A_p| + 1 = |A| - p + 1$. On a alors $p = 0$ et G est sans cycle.

On vient de voir que si G vérifie (iii) alors il vérifie (ii). On en déduit qu'entre deux sommets il existe un chemin puisque G est connexe, et ce chemin est unique car G est sans cycle, d'où (iii) implique (iv).

Supposons que G vérifie (iv), G est alors connexe. Soit uv une arête de G , si le graphe partiel de G induit par $A \setminus \{uv\}$ est connexe, il existe un autre chemin reliant u à v , ce qui contredit (iv). On a bien (iv) implique (v).

Enfin, si G vérifie (v) et si G possède un cycle $u = u_0u_1u_2 \dots u_{l-1}u_l = u$, le graphe partiel de G induit par $A \setminus \{u_{l-1}u_l\}$ est connexe ce qui contredit (v), d'où (v) implique (i). \square

Remarque. — On vient de voir que tout graphe connexe admet un graphe partiel qui est un arbre.

Définition 5.4. — Un arbre enraciné est un couple (H, r) formé d'un arbre H et d'un sommet r de H appelé racine.

Proposition 5.5. — Soit (H, r) un arbre enraciné et soient s et s' deux sommets de H . On pose $s \leq s'$ si et seulement si s' est sur le chemin reliant s à r . Cette relation est une relation d'ordre sur l'ensemble des sommets de H . Cet ordre admet r pour plus grand élément et les feuilles de H sont les éléments minimaux.

Remarque. — Si r est une feuille de l'arbre, il ne l'est plus pour l'arbre enraciné.

Dans un arbre enraciné, on dit que l'arête uv est issue du sommet u lorsque $v \leq u$. Dans ce cas, u est prédécesseur de v , et v un successeur de u .

5.2. X-arbre. — Nous allons définir la notion de X -arbre.

Définition 5.6. — Soit X un ensemble fini. Un X -arbre est un couple (H, f) , formé d'un arbre $H = (S, A)$ et d'une fonction f de X dans S telle que pour tout $v \in S \setminus f(X)$, $d(v) \geq 3$. La fonction f est l'étiquetage du X -arbre. Les sommets dans $f(X)$ sont appelés sommets réels, les sommets dans $S \setminus f(X)$ sont appelés sommets latents.

Deux X -arbres (H, f) et (H', f') sont isomorphes si et seulement si il existe une bijection g de S dans S' telle que st est une arête de A si et seulement si $g(s)g(t)$ est une arête de A' et $g \circ f = f'$.

Définition 5.7. — *Suivant les propriétés de la fonction f , on distingue plusieurs types de X -arbres.*

- Si $f(X)$ est l'ensemble des feuilles de H , on dit que le X -arbre (H, f) est libre.
- Si f est surjective, autrement dit si $S = f(X)$, on dit que (H, f) est contraint.
- Si f est injective, on dit que (H, f) est séparé.

Définition 5.8. — *Notons Y l'union de l'ensemble fini X et d'un cimetière $\{\partial\}$. Un X -arbre hiérarchique est un triplet (H, f, r) tel que :*

- (H, r) est un arbre enraciné ;
- (H, f) est un Y -arbre séparé ;
- $f(X)$ est l'ensemble des feuilles de H ;
- $f(\partial) = r$

Avec cette définition, on ne distingue pas l'ensemble des feuilles d'un X -arbre hiérarchique de l'ensemble X . On peut donc omettre la fonction f et noter un X -arbre hiérarchique comme un arbre enraciné (H, r) .

En utilisant la relation d'ordre entre les sommets d'un arbre enraciné, on peut définir la classe associée au sommet s de l'arbre hiérarchique (H, r) . C'est l'ensemble des feuilles situées « en dessous » du sommet s . Formellement cela donne

Définition 5.9. — *La classe associée au sommet s de l'arbre hiérarchique (H, r) est l'ensemble*

$$C(s) = \{x \in X, x \leq s\}.$$

En particulier, on a $C(r) = X$ et pour tout $x \in X$, $C(x) = \{x\}$. Si $s \leq s'$, on a l'inclusion $C(s) \subset C(s')$. Enfin, si s et s' ne sont pas comparables, $C(s) \cap C(s') = \emptyset$.

5.3. X -arbre valué ou dendrogramme. — Nous touchons enfin au but, définir ce qu'est un arbre phylogénétique.

Définition 5.10. — *Un X -arbre valué est un triplet (H, f, L) où (H, f) est un X -arbre et (H, L) est un arbre valué. On définit de même un arbre hiérarchique valué comme un triplet (H, r, L) où (H, r) est un arbre hiérarchique et où (H, L) est un arbre valué.*

Dans le cas des arbres hiérarchiques, on va utiliser la notion de valuation pour traduire l'idée de « niveaux de formation » des classes. Par exemple dans les arbres représentant l'évolution d'une séquence ou d'une espèce, la valuation représente le temps. En particulier pour une collection de séquences actuelles, ceci n'a de sens que si les chemins reliant les feuilles à la racine, qui représente la séquence ancestrale, ont la même longueur.

Définition 5.11. — *Un dendrogramme est un arbre hiérarchique valué tel que tous les chemins des feuilles à la racine ont même longueur.*

PARTIE III RECONSTRUCTION DE L'ARBRE PHYLOGÉNÉTIQUE DE SÉQUENCES D'ADN

Rappelons notre problème. On suppose que l'on a à notre disposition une collection de séquences d'ADN issues d'une même séquence ancestrale. De plus, on a été

capable de mettre une distance pertinente entre chaque couple de séquences de notre collection. Est-il possible de reconstruire l'arbre phylogénétique de ces séquences ? Si oui, et c'est ce qui intéresse les biologistes, avec quel algorithme, quelle vitesse et quelle robustesse ? Ce sont des questions auxquelles nous essaierons de répondre dans les mois à venir.

6. Les mesures de proximité et leurs représentations

6.1. Écarts et distances. —

Définition 6.1. — *Un indice d'écart sur l'ensemble X est une fonction δ , à valeurs réelles, définie sur $X \times X$ et vérifiant les deux conditions ci-dessous :*

- pour tout $x, y \in X$, $\delta(x, y) = \delta(y, x)$, (symétrie)
- pour tout $x, y \in X$, $\delta(x, y) \geq 0$ et $\delta(x, x) = 0$, (positivité).

Un espace écarté est un couple (X, δ) formé d'un ensemble X et d'un indice d'écart δ sur X . Une isométrie d'un espace écarté (X, δ) vers un espace écarté (X', δ') est une fonction f de X dans X' telle que pour tout $x, y \in X$, $\delta(x, y) = \delta'(f(x), f(y))$.

Définition 6.2. — *Un indice de distance sur X est un indice d'écart δ sur X tel que pour tout $x, y \in X$, $\delta(x, y) = 0$ entraîne $x = y$.*

Lemme 6.3. — *Soit f une isométrie de l'espace écarté (X, δ) vers (X', δ') . Si δ est un indice de distance alors f est injective. Si δ' est un indice de distance et f est injective alors δ est un indice de distance.*

Démonstration. — Soient $x, y \in X$ tels que $f(x) = f(y)$, on a alors $\delta(x, y) = \delta'(f(x), f(y)) = 0$. Puisque δ est un indice de distance, on a alors $x = y$. D'autre part soient $x, y \in X$ tels que $\delta(x, y) = 0$, on a alors $\delta'(f(x), f(y)) = 0$. Si δ' est un indice de distance, on a alors $f(x) = f(y)$ et si f est injective, $x = y$. \square

Définition 6.4. — *On appelle écart (respectivement distance), un indice d'écart δ (respectivement un indice de distance) qui vérifie l'inégalité triangulaire, pour tout $x, y, z \in X$, $\delta(x, z) \leq \delta(x, y) + \delta(y, z)$. Un espace métrique est un couple (X, δ) où δ est une distance sur X .*

Tout ce qui précède est un rappel sur les espaces métriques. En analyse de données on ne voit pas trop à quoi peut servir l'inégalité triangulaire, mais rappelons que notre but est de travailler sur des représentations de nos données, ici des arbres, et c'est la géométrie de ces arbres qui nécessite l'inégalité triangulaire. Donnons à présent une définition formelle à la notion de représentation. Soit \mathbf{E} un ensemble d'espaces écartés.

Définition 6.5. — *on dit que l'espace écarté (X, δ) admet une \mathbf{E} -représentation, ou qu'il est représentable dans \mathbf{E} , si et seulement si il existe $(S, d) \in \mathbf{E}$ et une isométrie f de (X, δ) vers (S, d) . Le triplet (S, d, f) est alors appelé une \mathbf{E} -représentation de (X, δ) et (S, d) est l'espace de représentation.*

Exemple. — \mathbf{E} n'a qu'un seul élément, la droite réelle \mathbf{R} , munie de sa distance canonique. une $\{\mathbf{R}\}$ -représentation de (X, δ) est appelée représentation linéaire de (X, δ) . c'est une fonction f de X dans \mathbf{R} telle que pour tout $x, y \in X$, $\delta(x, y) = |f(x) - f(y)|$.

Proposition 6.6. — *Un espace écarté (X, δ) admet une représentation linéaire si et seulement si l'indice d'écart δ vérifie pour tout $x, y, z \in X$:*

$$\max(\delta(x, y), \delta(x, z)) = \min(\delta(x, y) + \delta(y, z), \delta(x, z) + \delta(z, y)).$$

Démonstration. — L'inégalité ci-dessus signifie que

$$\delta(x, y) \leq \delta(x, z) \text{ entraîne } \delta(x, z) = \delta(x, y) + \delta(y, z).$$

Ceci est vérifié par la distance canonique de \mathbf{R} , la condition est donc nécessaire.

Réciproquement, considérons x_0 un élément de X et posons $f(x) = \delta(x_0, x)$. Soient $x, y \in X$, supposons $\delta(x_0, x) \leq \delta(x_0, y)$, il vient alors $\delta(x_0, y) = \delta(x_0, x) + \delta(x, y)$, c'est à dire $f(y) - f(x) = \delta(x, y)$. En considérant de même le cas $\delta(x_0, x) \geq \delta(x_0, y)$, on trouve $\delta(x, y) = |f(x) - f(y)|$. \square

Exemple. — \mathbf{E} est la classe de tous les espaces euclidiens, c'est à dire de tous les espaces vectoriels \mathbf{R}^p munis d'un produit scalaire. on dit qu'une distance δ sur X est euclidienne si et seulement si δ est représentable dans \mathbf{E} .

Proposition 6.7. — Une condition nécessaire et suffisante pour que la distance δ soit euclidienne est qu'il existe $x_0 \in X$ tel que la matrice de terme général :

$$A_{x_0}(x, y) = \frac{1}{2} (\delta(x_0, x)^2 + \delta(x_0, y)^2 - \delta(x, y)^2)$$

soit positive, c'est à dire pour tout $u \in \mathbf{R}^{|X|}$, $u^T A u \geq 0$. La dimension minimale d'un espace de représentation est égale au rang de la matrice A . De plus cette propriété est indépendante du choix de x_0 .

Démonstration. — Montrons d'abord que la condition est nécessaire. Supposons que δ est euclidienne, une représentation euclidienne f de (X, δ) étant injective, on peut identifier x et $f(x)$. Notons alors x^i , pour $1 \leq i \leq p$, les p coordonnées de x dans \mathbf{R}^p . Il vient alors,

$$\delta(x, y)^2 = \sum_{i=1}^p (x^i - y^i)^2.$$

Soit x_0 un élément de X , nous avons

$$A_{x_0}(x, y) = 2 \sum_{i=1}^p (x_0^i - x^i)(x_0^i - y^i) = 2(x_0 - x)^T (x_0 - y).$$

La matrice A_{x_0} est donc la matrice de Gram des $|X|$ vecteurs $x_0 - x$ de \mathbf{R}^p . Soit u un élément de $\mathbf{R}^{|X|}$. Il vient

$$\begin{aligned} u^T A_{x_0} u &= \sum_{x \in X} (A_{x_0} u)(x) u(x) \\ &= \sum_{x, y \in X} A_{x_0}(x, y) u(y) u(x) \\ &= 2 \sum_{x, y \in X} (u(x)(x_0 - x))^T (u(y)(x_0 - y)) \\ &= 2 \left(\sum_{x \in X} u(x)(x_0 - x) \right)^T \left(\sum_{y \in X} u(y)(x_0 - y) \right) \\ &= 2 \left\| \sum_{x \in X} u(x)(x_0 - x) \right\|^2. \end{aligned}$$

Pour tout x_0 , la matrice A_{x_0} est positive. De plus le rang de A_{x_0} est égal à la dimension de l'espace vectoriel engendré par la famille $\{x_0 - x, x \in X\}$, encore égal à la dimension de $\text{vect}(x, x \in X)$ moins 1.

Montrons maintenant que la condition est suffisante. Soit x_0 un élément de X tel que la matrice A_{x_0} soit positive et de rang p . Il existe une matrice B de taille $n \times p$

de rang p , telle que $A_{x_0} = BB^T$. De la relation

$$A_{x_0}(x, y) = \frac{1}{2} (\delta(x_0, x)^2 + \delta(x_0, y)^2 - \delta(x, y)^2),$$

il vient

$$\delta(x, y)^2 = A_{x_0}(x, x) + A_{x_0}(y, y) - 2A_{x_0}(x, y) = \sum_{t \in X} (B(x, t) - B(y, t))^2.$$

Posons alors $f(x) = (B(x, t), t \in X)$, on obtient une représentation euclidienne de (X, δ) . \square

6.2. Distances arborées. — Soit (H, L) un arbre valué, Rappelons que l'on désigne par S l'ensemble de ces sommets, par A l'ensemble de ces arêtes, que pour st une arête de H , $L(st)$ est la longueur de l'arête st et que pour un chemin de H , la longueur de ce chemin est la somme des longueurs des arêtes qui le composent.

Définition 6.8. — On définit alors la distance additive d_L , notée simplement d quand il n'y a pas d'ambiguïté par : pour tout $s, t \in S$, $d_L(s, t)$ vaut la longueur du chemin entre s et t . La fonction d est une distance au sens de la définition 6.4.

Lemme 6.9. — Soit (H, L) un arbre valué, soit (S, d) l'espace métrique associé et soit $s, t \in S$. une condition nécessaire et suffisante pour que st soit une arête de H est qu'il n'existe aucun sommet u de H , distinct de s et t , tel que $d(s, t) = d(s, u) + d(u, t)$.

Démonstration. — La condition est nécessaire. Si l'on a $d(s, t) = d(s, u) + d(u, t)$, il existe un chemin qui passe par u reliant s à t et st n'est donc pas une arête de H , car H est un arbre.

La condition est suffisante car si st n'est pas une arête de H , il existe un chemin non réduit à une arête reliant s à t et il suffit de choisir un des sommets de ce chemin autre que les extrémités. \square

Lemme 6.10. — Soit (H, L) et (H', L') deux arbres valués dont les espaces métriques associés (S, d) et (S', d') coïncident. Alors $(H, L) = (H', L')$.

Démonstration. — Supposons que (S, d) et (S', d') coïncident, en appliquant le lemme 6.9, on voit que st est une arête de H si et seulement si st est une arête de H' donc $H = H'$. De plus, $L(st) = d(s, t) = d'(s, t) = L'(st)$, donc $L = L'$. \square

Compte tenu du lemme 6.10, on peut identifier la classe des arbres valués (H, L) et la classe (S, d) des espaces métriques qui leur sont associés. Notons **ARB** cette classe. On peut voir que tout indice d'écart représentable dans **ARB** vérifie l'inégalité triangulaire, c'est donc un écart.

Définition 6.11. — On appelle écart arboré (distance arborée), tout indice d'écart (tout indice de distance) représentable dans **ARB**.

Nous allons aboutir à une condition nécessaire et suffisante pour qu'un écart sur soit arboré, la condition des quatre points.

CONCLUSION

Ce travail est le point de départ d'une thèse. Récapitulons les différentes questions auxquelles nous allons essayer de répondre. Nous devons trouver un estimateur consistant de la distance séparant une séquence actuelle d'une séquence ancestrale.

Ensuite, il nous faut déterminer un estimateur de la distance séparant deux séquences actuelles que l'on sait issues d'une séquence ancestrale. À partir de cet estimateur, nous devons déterminer un algorithme pour établir la phylogénie d'une collection de séquences actuelles.

Références

- [1] J.-P. BARTHÉLÉMY & A. GUÉNOCHE – *Les arbres et les représentations de proximité*, Masson, 1988.
- [2] J. BÉRARD, J.-B. GOUÉRÉ & D. PIAU – « Solvable models of neighbor-dependent nucleotide substitution processes », *e-arxiv math.PR/0510034* (2005).
- [3] L. DURET & N. GALTIER – « The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact », *Molecular biology and evolution* **17** (2000), p. 1620–1625.
- [4] Z. YANG – *Computational Molecular Evolution*, Oxford Series in Ecology and Evolution, 2006.

15 juin 2007

MIKAEL FALCONNET