Estimations de temps évolutifs pour des modèles de substitutions avec interaction entre les sites

Mikael Falconnet

Institut Fourier - Université de Grenoble

Mai 2010

Colloque Jeunes Probabilistes et Statisticiens

Phylogénie

Définition (Wikipedia)

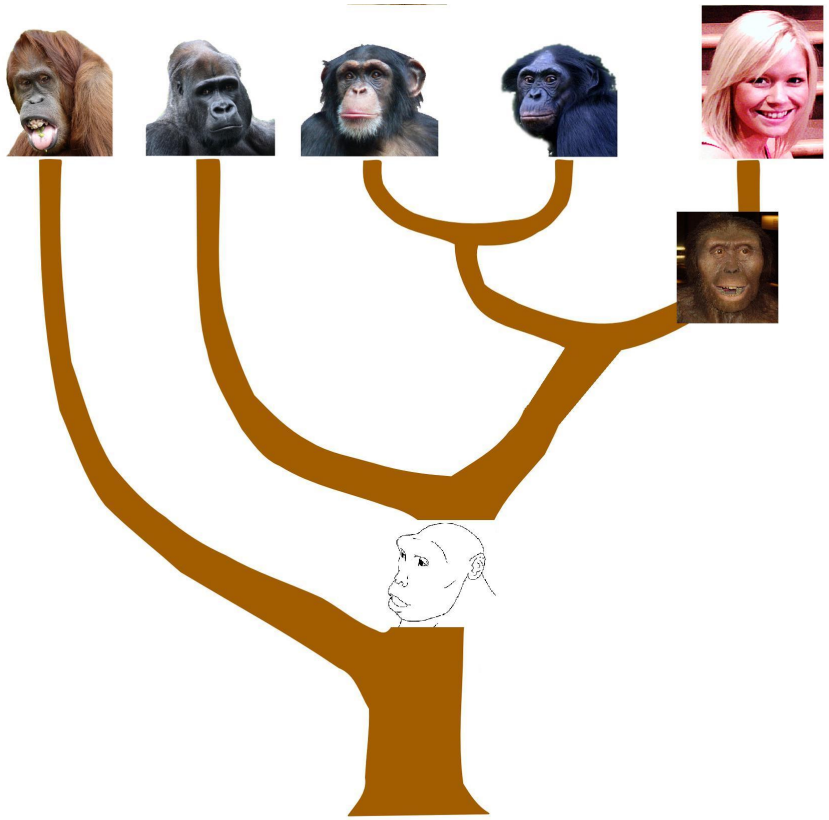
En biologie, la phylogénie est l'étude de la formation et de l'évolution des organismes vivants en vue d'établir leur parenté.

Idée

Le degré de ressemblance est corrélé au degré de parenté.

Arbre phylogénétique

L'évolution est vue comme un processus de branchement, dans lequel les populations changent au cours du temps et peuvent diverger en différents groupes. On visualise cette évolution dans un arbre phylogénétique.

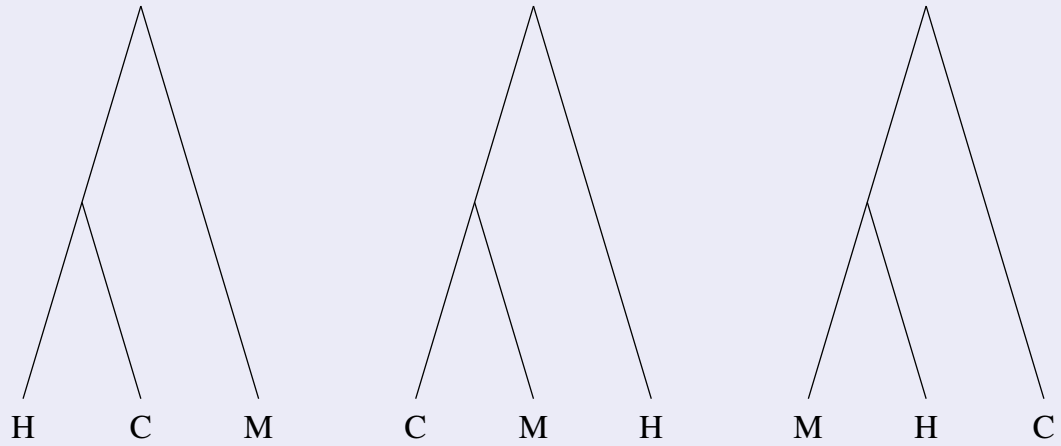


Arbre phylogénétique

Le but de la phylogénie est de construire l'arbre phylogénétique le plus vraisemblable entre des organismes actuels.

Premier problème : la topologie de l'arbre

Parmis les arbres suivants, lequel représente l'arbre phylogénétique des trois espèces suivantes : Homme, Chimpanzée, Macaque?

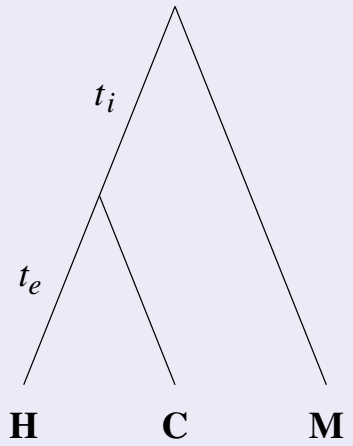


Arbre phylogénétique

En supposant que l'on connaisse la topologie de l'arbre, il vient une autre question.

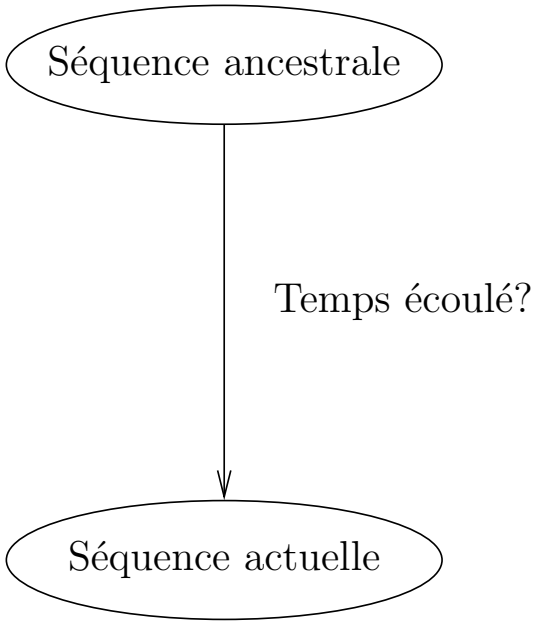
Second problème : les longueurs des branches

Quelles sont les longueurs des branches (t_i, t_e) dans l'arbre phylogénétique de l'Homme, du Chimpanzée et du Macaque?



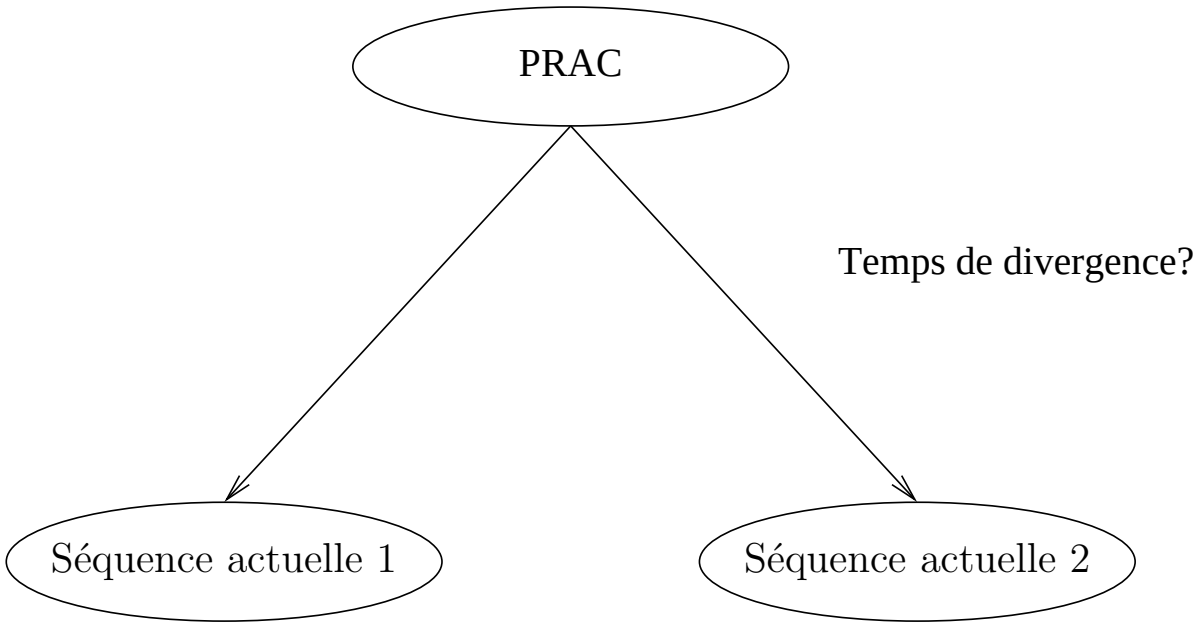
Premier problème

On suppose qu'une séquence d'ADN actuelle est issue d'une séquence d'ADN ancestrale. Le but est d'estimer le temps écoulé entre les deux séquences.

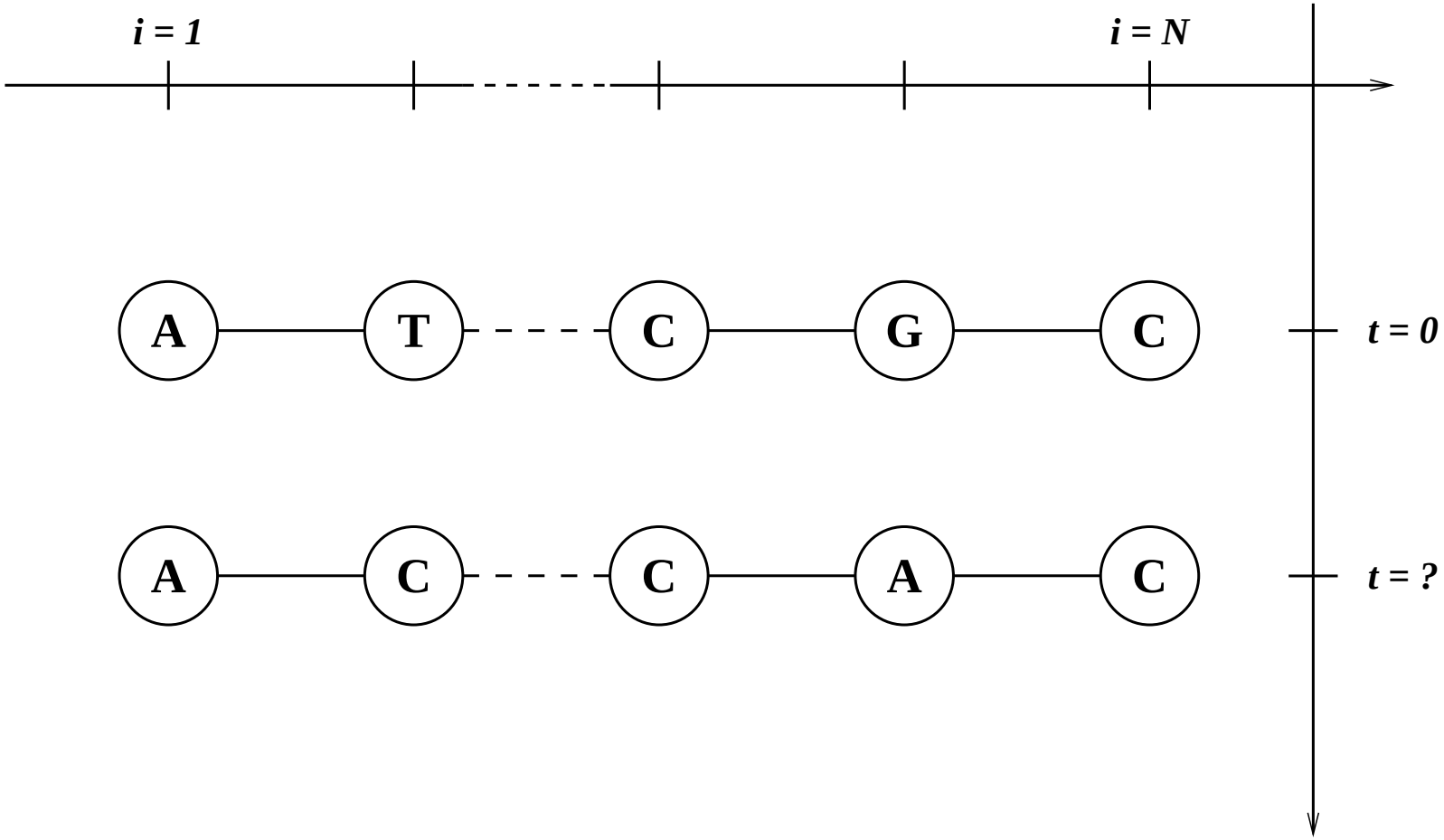


Second problème

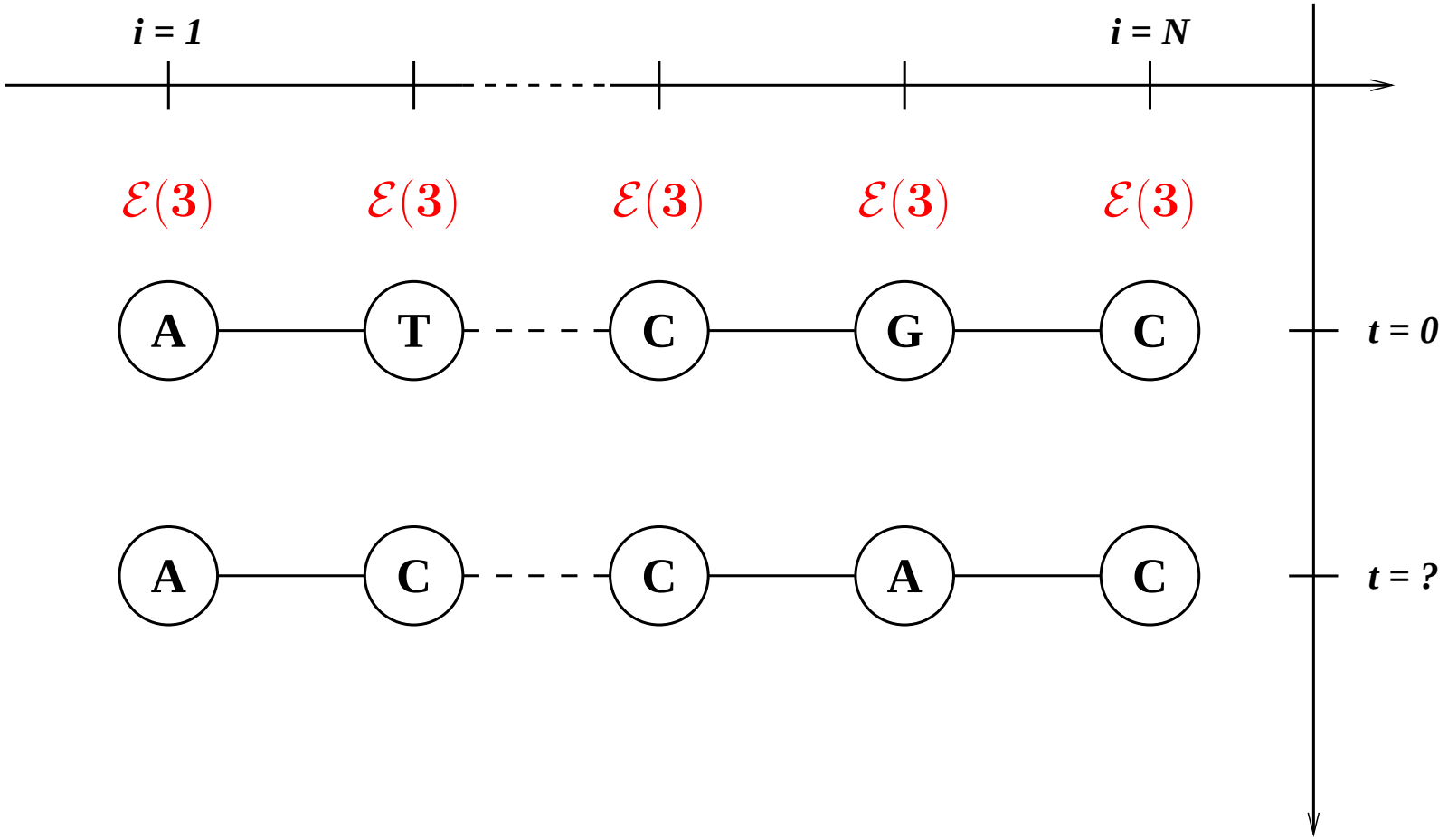
Bien sûr, on ne dispose pas, en général, d'une séquence ancestrale, on doit travailler avec des séquences actuelles. En supposant que deux séquences d'ADN actuelles sont issues d'une séquence ancestrale commune, on veut estimer le temps depuis lequel ces deux séquences ont divergé.



Heuristique



Heuristique



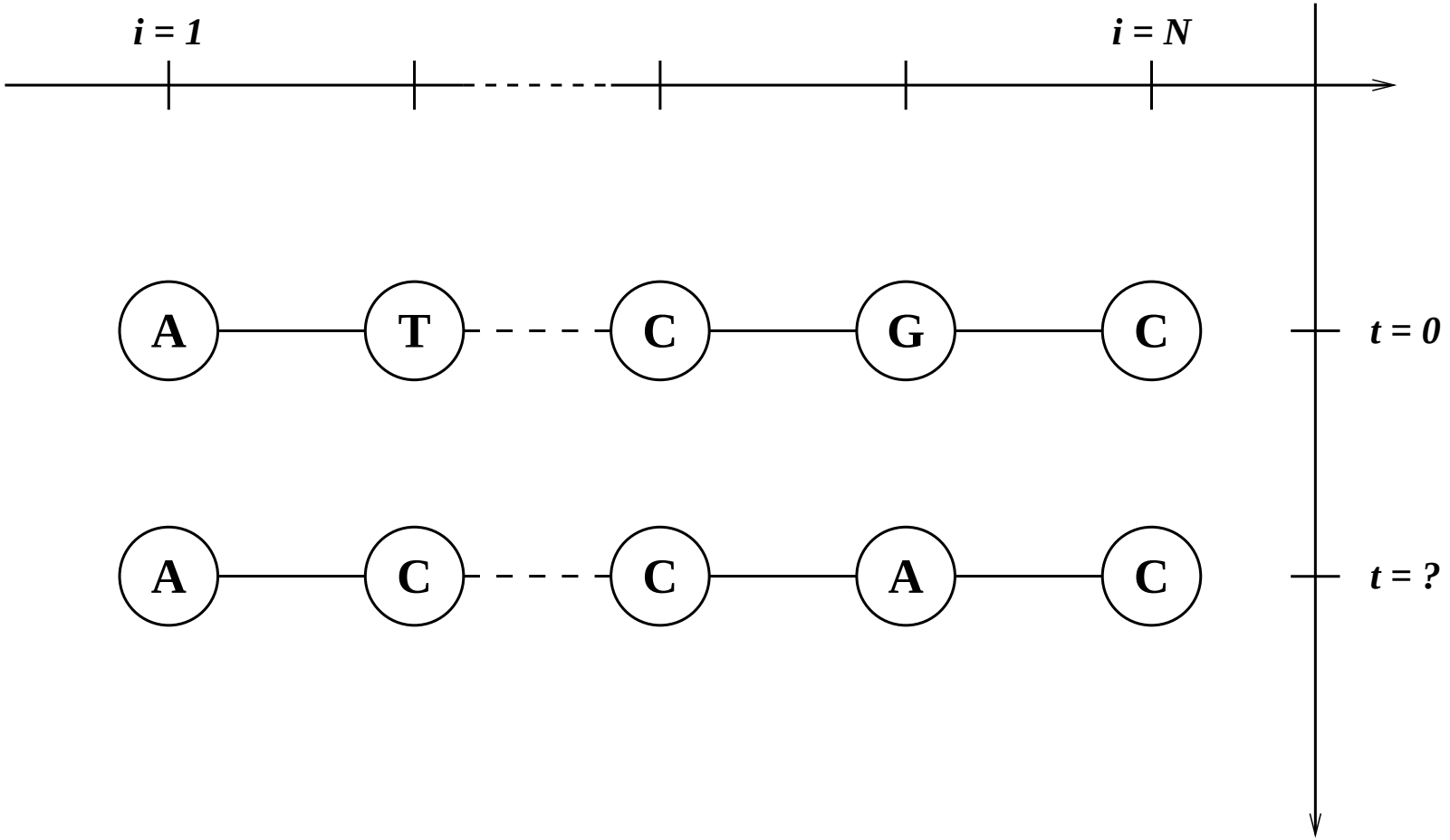
Le modèle de Jukes et Cantor

Définition

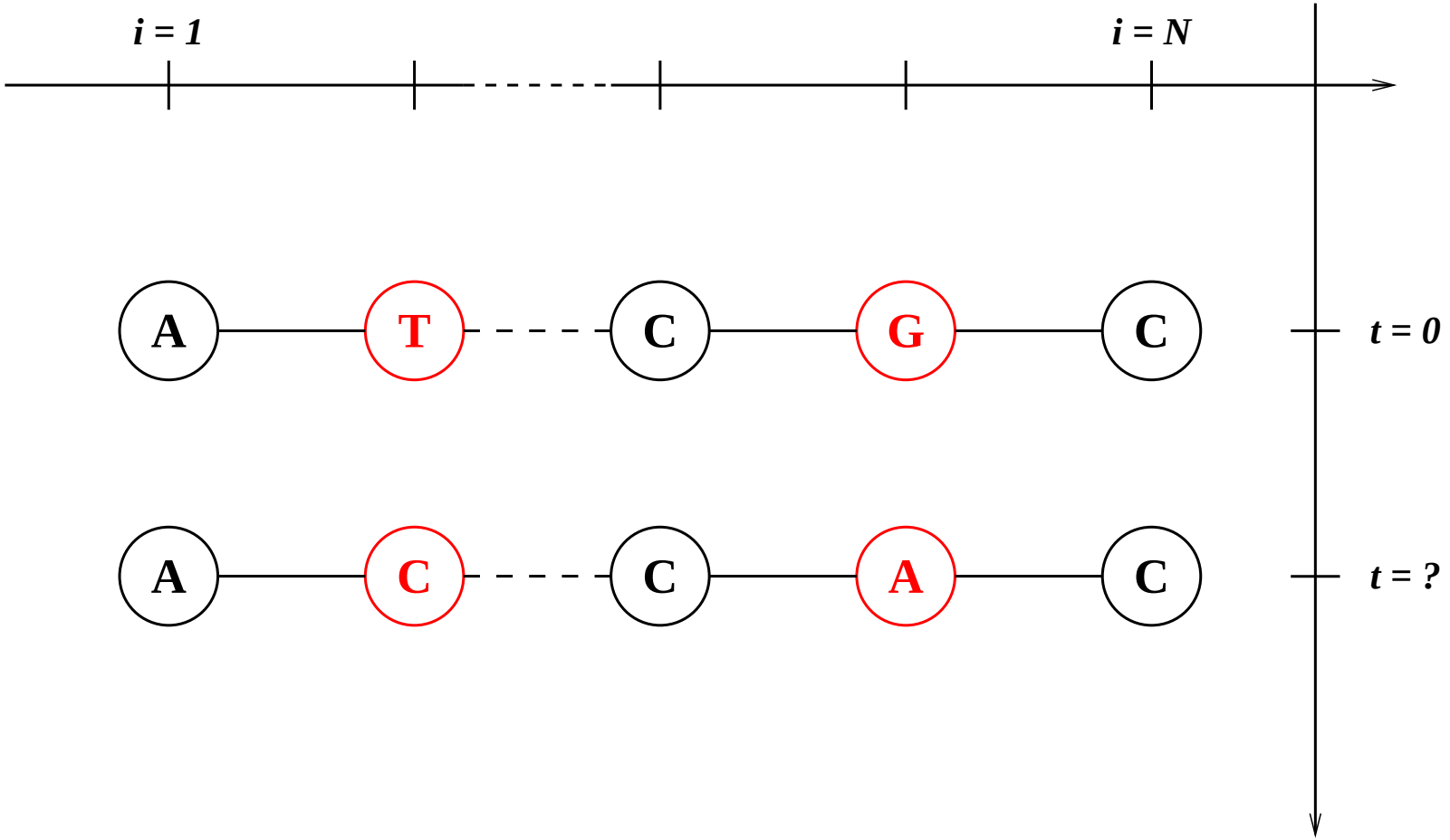
Le modèle de Jukes et Cantor est un processus de Markov, noté $(X_{1:N}(t))_{t\geqslant 0}$, *sur* $\mathscr{A}^N$, *où* $\mathscr{A} = \{A, T, C, G\}$ *et* $N \in \mathbb{N}^\times$. *Chaque site évolue indépendamment des autres en suivant le générateur infinitésimal suivant :*

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & 1 & 1 & 1 \\ 1 & \cdot & 1 & 1 \\ 1 & 1 & \cdot & 1 \\ 1 & 1 & 1 & \cdot \end{pmatrix} \end{array}.$$

Le modèle de Jukes et Cantor



Le modèle de Jukes et Cantor



Le modèle de Jukes et Cantor

Définition

On note Q_{obs} la quantité observée définie par

$$Q_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i(t) \neq X_i(0)\}.$$

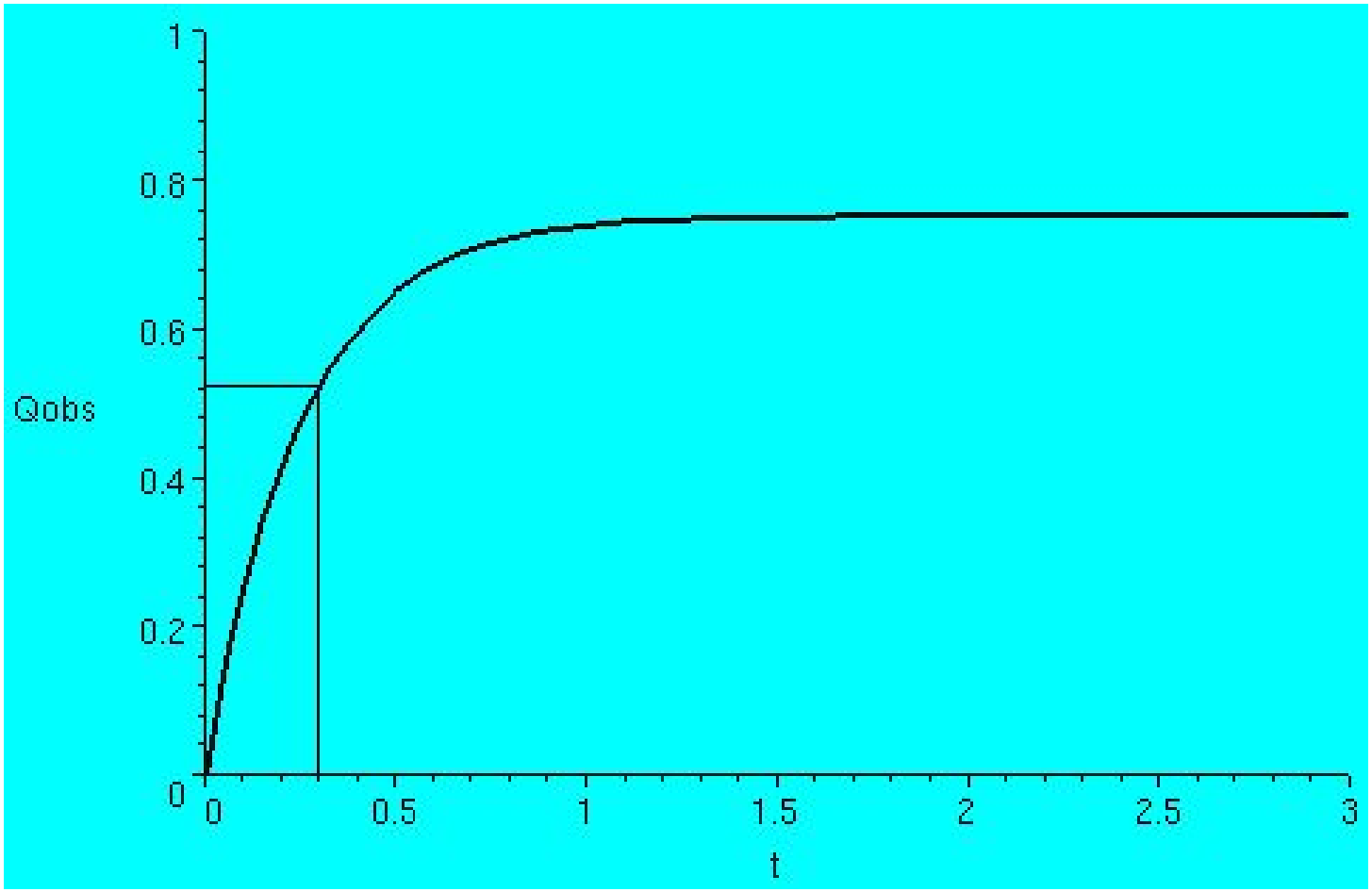
Les v.a. $(\mathbf{1}\{X_i(t) \neq X_i(0)\})_{i=1}^{N}$ sont des v.a. de Bernoulli i.i.d., de paramètre $q(t)$ donné par

Proposition

Pour tout $1 \leqslant i \leqslant N$, pour tout t positif

$$q(t) = \mathbb{P}(X_i(t) \neq X_i(0)) = \frac{3}{4}(1 - \mathrm{e}^{-4t}).$$

Représentation de $t \mapsto q(t)$



Estimateur consistant du temps écoulé

Définition

On note T l'estimateur du temps écoulé défini comme l'unique solution en τ de l'équation

$$Q_{\text{obs}} = q(\tau).$$

Consistance

Comme

$$Q_{\text{obs}} \xrightarrow[N\to+\infty]{p.s.} q(t),$$

on en déduit que

Proposition

$$T \xrightarrow[N\to+\infty]{p.s.} t.$$

Intervalle de confiance

Comme $(\mathbf{1}\{X_i(t) \neq X_i(0)\})_i$ est une suite de v.a. i.i.d. suivant une loi de Bernoulli de paramètre $q(t)$, il vient

TCL pour Q_{obs}

$$\sqrt{N}\,(Q_{\text{obs}} - q(t)) \xrightarrow[N\to+\infty]{Loi} \mathcal{N}(0, \sigma^2(t)),$$

où $\sigma^2(t) = q(t)(1 - q(t))$.

Intervalle de confiance

Méthode delta

Supposons que la suite de variables aléatoires (X_n) vérifie

$$\sqrt{n}(X_n - \theta) \xrightarrow[N\to+\infty]{Loi} \mathcal{N}(0, \sigma^2),$$

alors dès que $\varphi'(\theta)$ existe et est différente de zéro

$$\sqrt{n}(\varphi(X_n) - \varphi(\theta)) \xrightarrow[N\to+\infty]{Loi} \mathcal{N}(0, \sigma^2[\varphi'(\theta)]^2).$$

Intervalle de confiance

On applique la méthode delta à la fonction q^{-1} et on obtient le résultat suivant.

Proposition

$$\sqrt{N}\,(T - t) \xrightarrow[N\to+\infty]{Loi} \mathcal{N}\left(0, \frac{\sigma^2(t)}{q'(t)^2}\right).$$

Intervalle de confiance

A priori, pour construire un intervalle de confiance pour t à partir de la proposition précédente, on a besoin de connaître la valeur de $q'(t)$ et de $\sigma^2(t)$ qui dépendent de la quantité t qu'on cherche à estimer.

Lemme de Slutsky

Supposons que $X_n \xrightarrow[N\to+\infty]{Loi} X$ et $Y_n \xrightarrow[N\to+\infty]{\mathbb{P}} c$, alors

$$X_n + Y_n \xrightarrow[N\to+\infty]{Loi} X + c, \quad X_n Y_n \xrightarrow[N\to+\infty]{Loi} cX,$$

$$X_n / Y_n \xrightarrow[N\to+\infty]{Loi} X/c, \quad \text{si } c \neq 0.$$

Intervalle de confiance

On applique le lemme de Slutsky avec

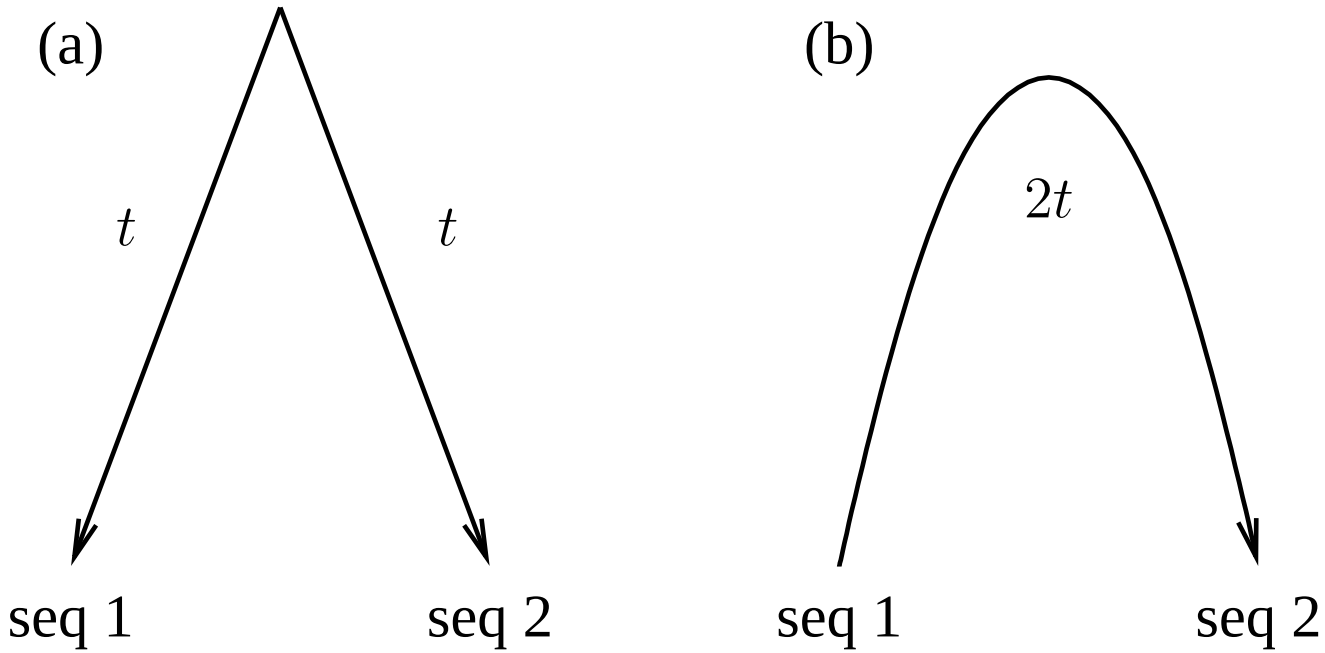
$$3 - 4Q_{\text{obs}} \xrightarrow[N\to+\infty]{p.s.} 3 - 4q(t) = q'(t),$$
$$Q_{\text{obs}}(1 - Q_{\text{obs}}) \xrightarrow[N\to+\infty]{p.s.} \sigma^2(t).$$

Théorème

$$(4Q_{\text{obs}} - 3)\sqrt{\frac{N}{Q_{\text{obs}}(1 - Q_{\text{obs}})}}\,(T - t) \xrightarrow[N\to+\infty]{Loi} \mathcal{N}(0, 1).$$

Estimateur consistant du temps de divergence

Comme le modèle de Jukes-Cantor est réversible, on a équivalence entre les deux schémas suivants



Estimateur consistant du temps de divergence

Proposition

Pour tout $1 \leqslant i \leqslant N$, *pour tout* t *positif ;*

$$\mathbb{P}\left(X_i^1(t) \neq X_i^2(t)\right) = q(2t).$$

Mécanismes du modèle de Jukes-Cantor avec influence CpG

Évolution indépendante

Le premier mécanisme est une évolution indépendante des sites comme dans le modèle de Jukes et Cantor. Le taux de substitution de x par y est de 1 pour touts nucléotides $x \neq y$ dans $\mathscr{A}$.

$$x \xrightarrow{1} y, \quad \forall x \neq y.$$

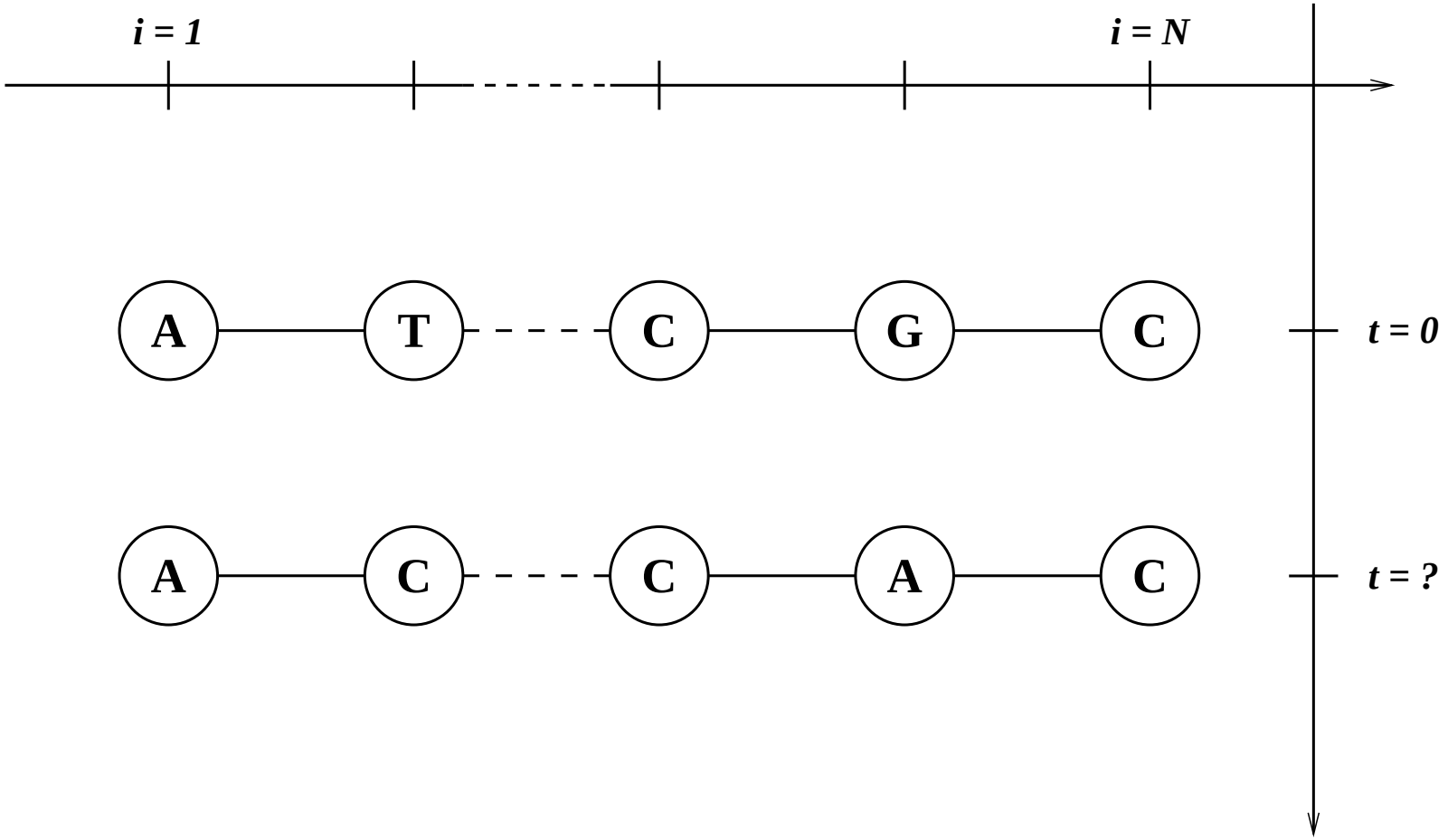
Mécanismes du modèle de Jukes-Cantor avec influence CpG

Influence du voisinage

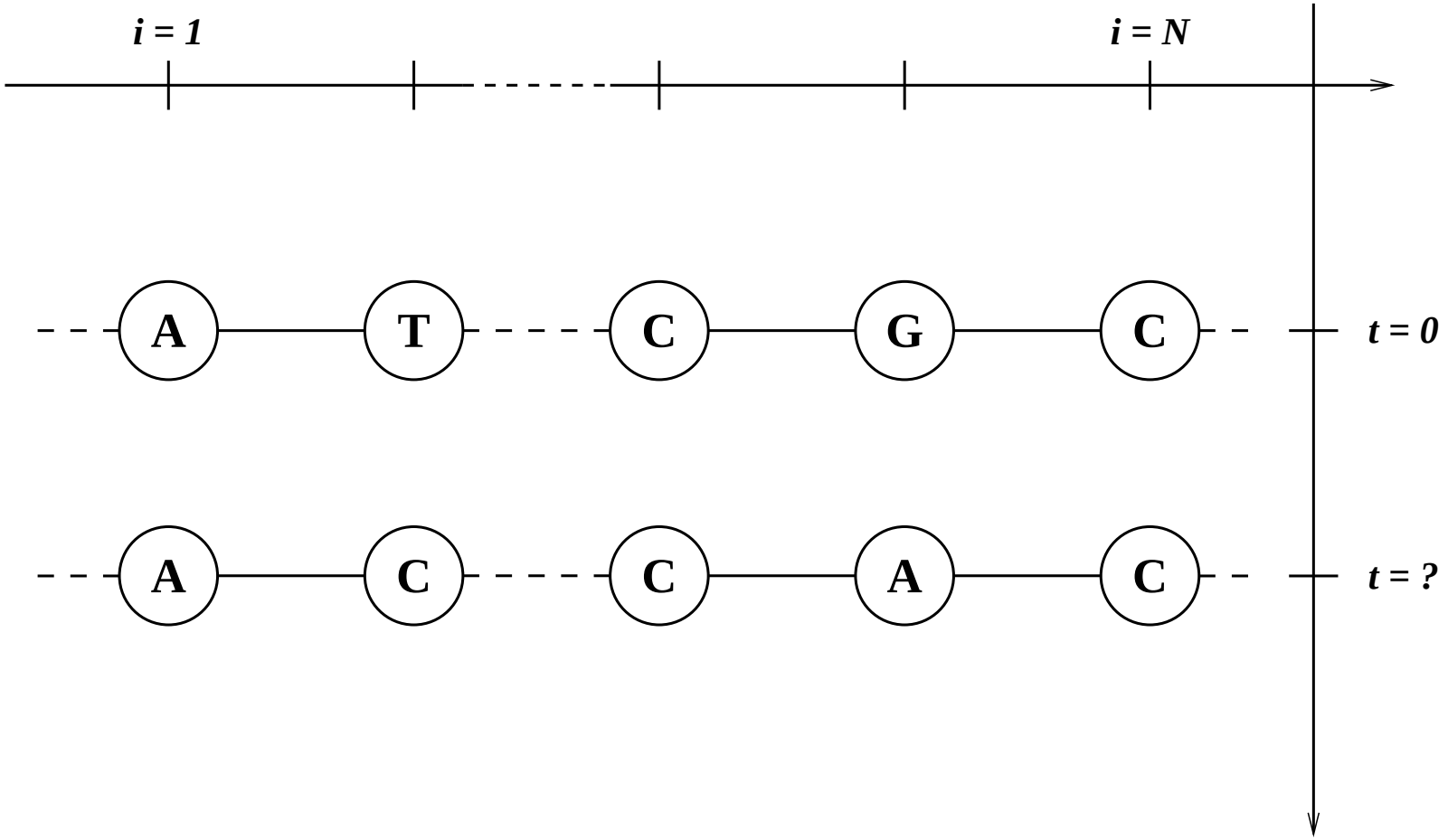
Un second mécanisme est ajouté, qui décrit les substitutions dues à l'influence du voisinage : on suppose que les taux de substitutions de la cytosine par la thymine et de la guanine par l'adenine sont augmentés de r dans les dinucléotides CpG.

$$CG \xrightarrow{r} TG \quad \text{and} \quad CG \xrightarrow{r} CA.$$

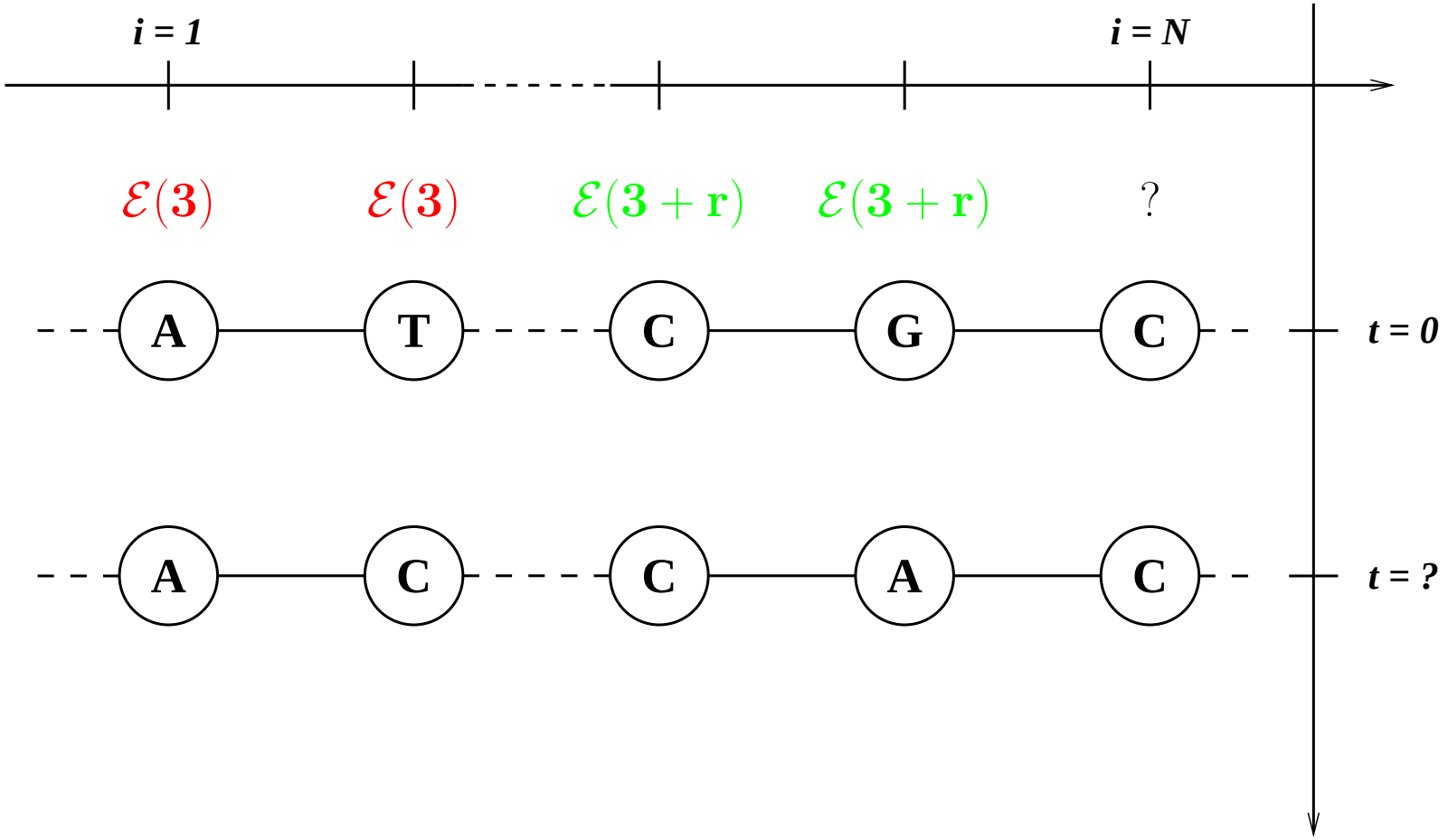
Heuristique



Heuristique

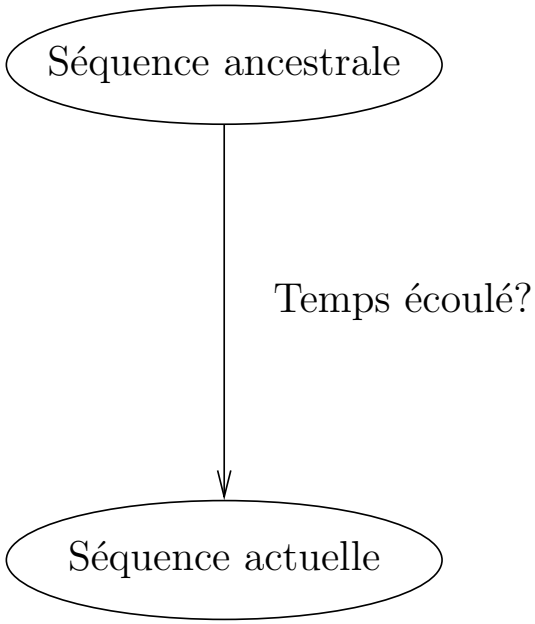


Heuristique

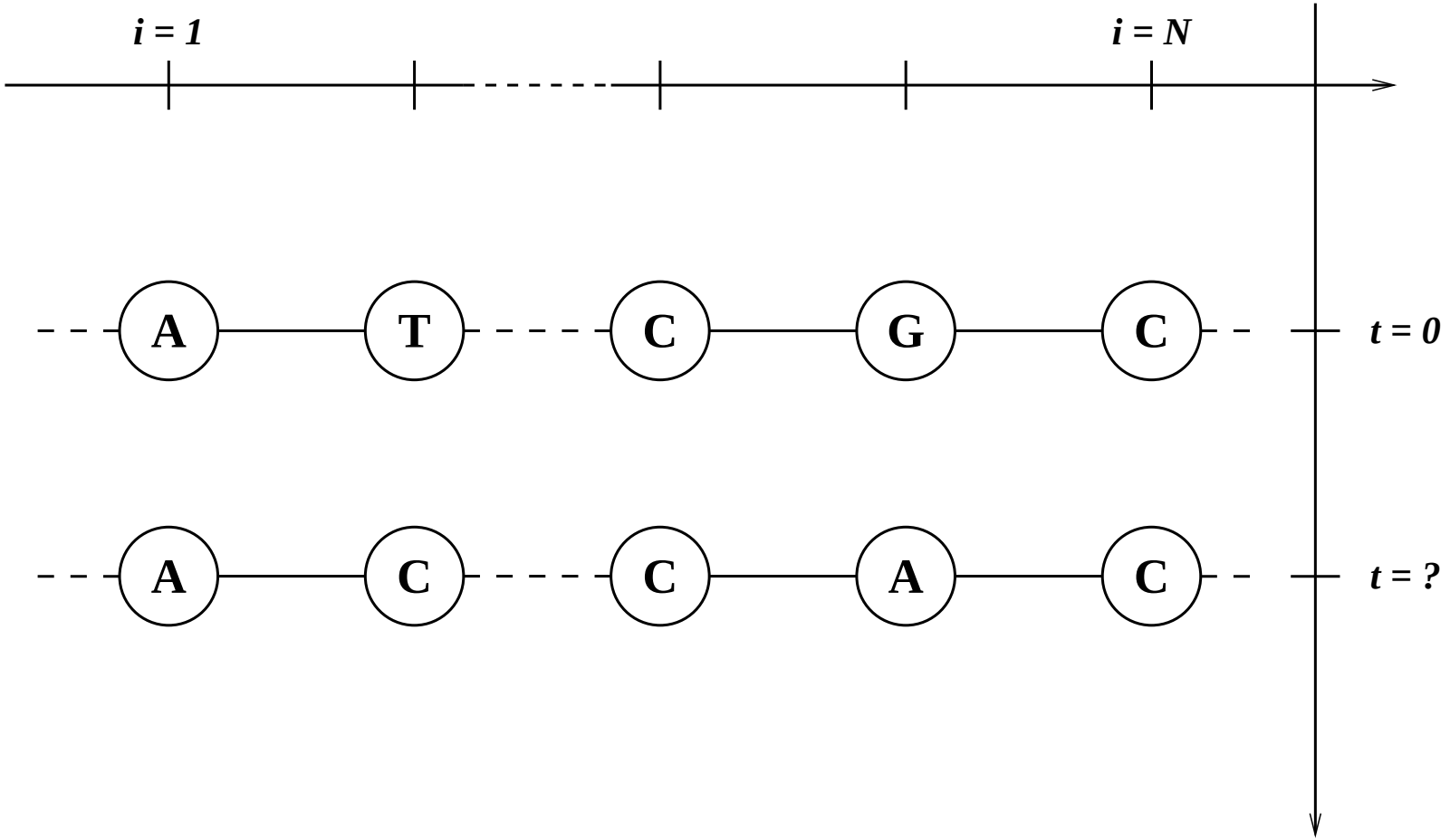


Premier problème

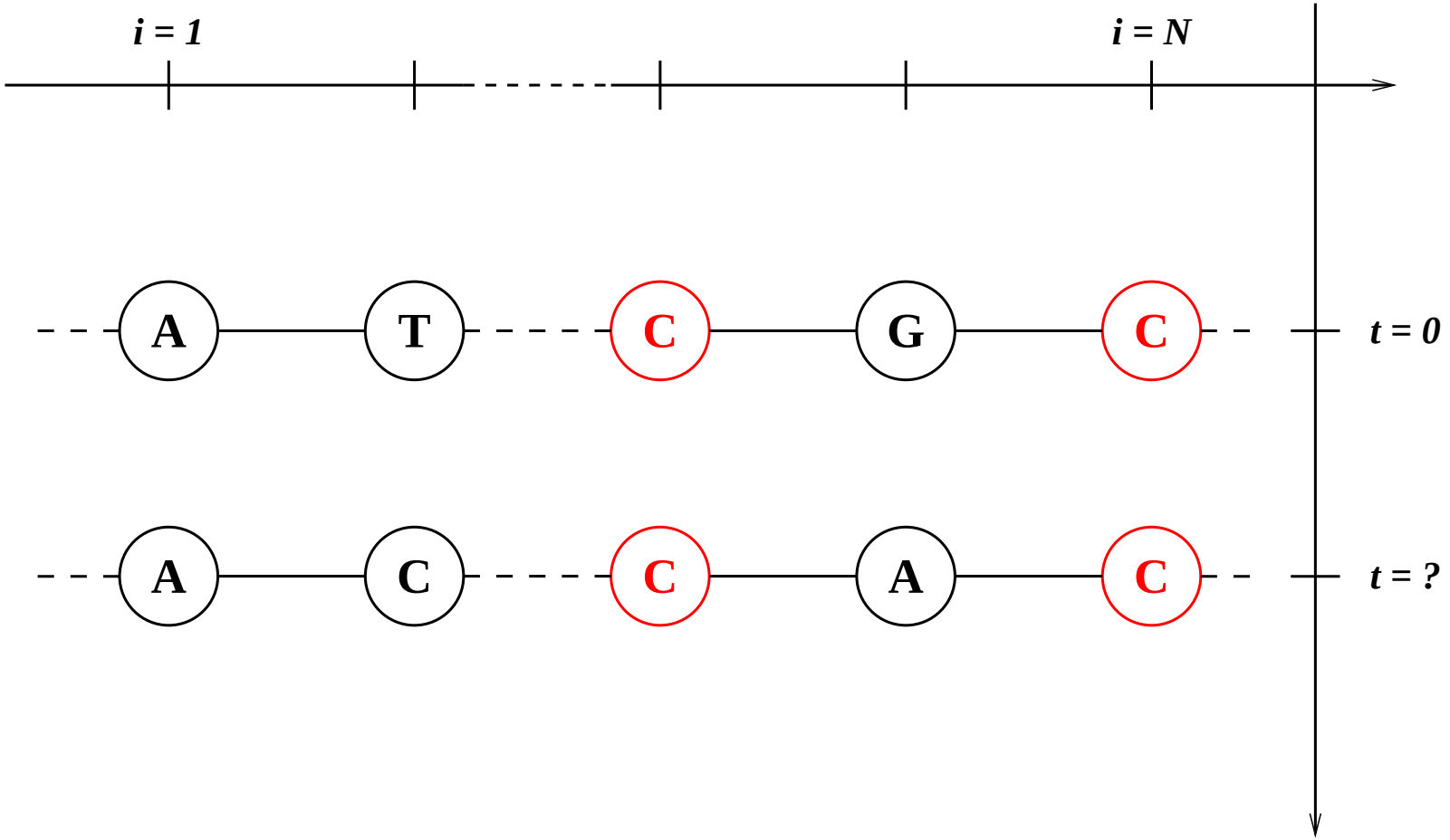
On suppose qu'une séquence d'ADN actuelle est issue d'une séquence d'ADN ancestrale. Le but est d'estimer le temps écoulé entre les deux séquences.



Idée



Idée



Notations et définitions

Définition

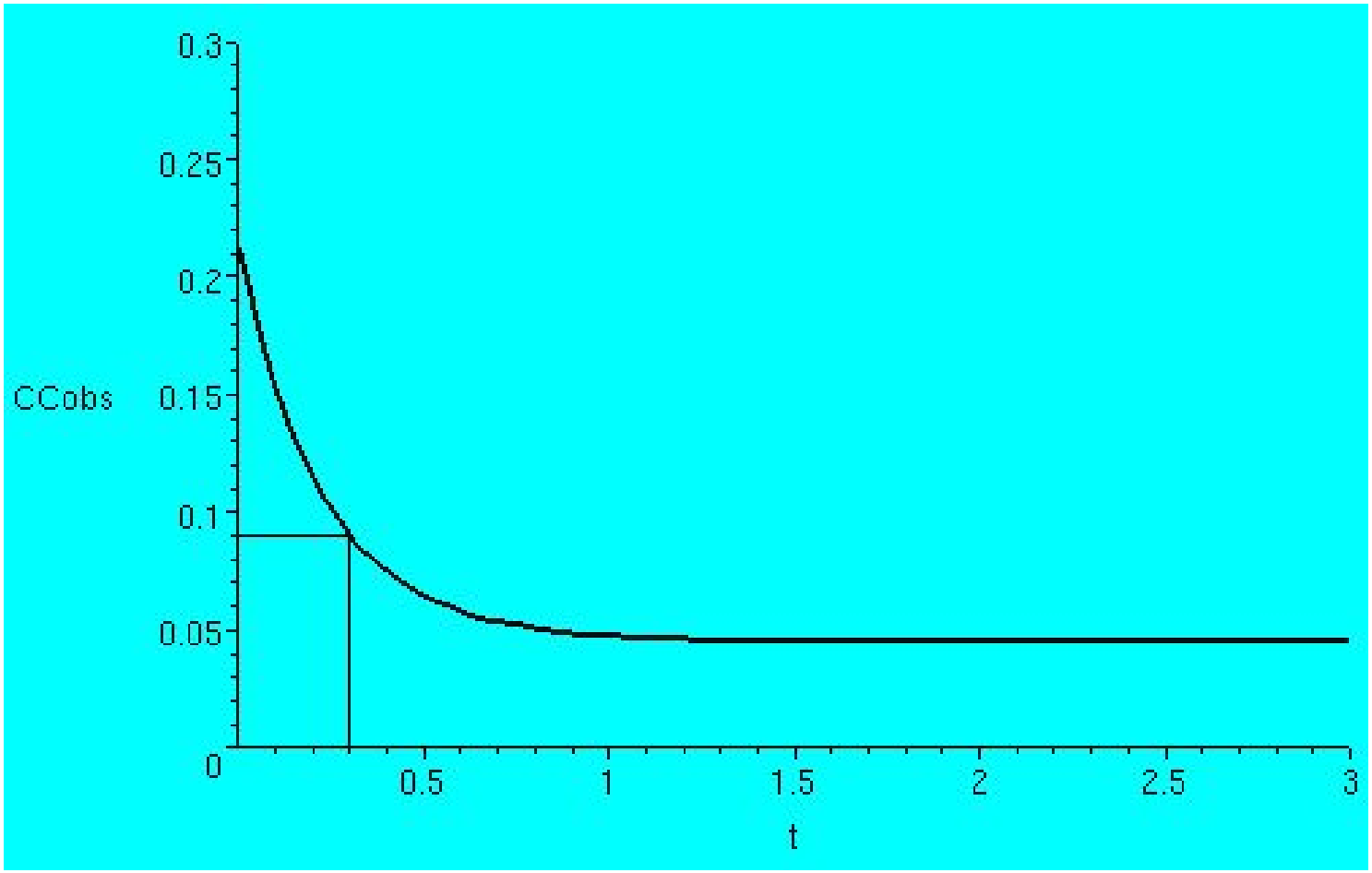
On note $(C, C)_{\text{obs}}$ *la quantitée observée définie par*

$$(C, C)_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i(t) = C, X_i(0) = C\}.$$

On note $(C, C)(t)$ *la fréquence des sites occupés par* C *aux temps* 0 *et* t*, c'est à dire*

$$(C, C)(t) = \lim_{N \to \infty} (C, C)_{\text{obs}}.$$

Représentation de $t \mapsto (C, C)(t)$



Notations et définitions

Définition

On note T_C l'estimateur du temps écoulé comme la solution en τ de l'équation

$$(C, C)(\tau) = (C, C)_{\text{obs}}.$$

On note κ^C_{obs} et ν^C_{obs} les quantités observées, définies par

$$\begin{aligned}
\kappa^C_{\text{obs}} &= 4(C, C)_{\text{obs}} + r(C*, CG)_{\text{obs}} - (C)_{\text{obs}}, \\
\nu^C_{\text{obs}} &= (C, C)_{\text{obs}} + 2(CC, CC)_{\text{obs}} + 2(C * C, C * C)_{\text{obs}} \\
&\quad - 5(C, C)^2_{\text{obs}}.
\end{aligned}$$

Résultat

Théorème (MF)

Supposons que la séquence ancestrale est à l'équilibre, alors dans le modèle JC + CpG,

$$\kappa_{\mathrm{obs}}^{C}\sqrt{N/\nu_{\mathrm{obs}}^{C}}(T_C - t) \xrightarrow[N\to+\infty]{Loi} \mathcal{N}(0,1).$$

Références I

J. Bérard, J.-B. Gouéré, and D. Piau.
Solvable models of neighbor-dependent nucleotide substitution processes.
Mathematical Biosciences, 211:56–88, 2008.

L. Duret and 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, 17:1620–1625, 2000.

M. Falconnet.
Phylogenetic distances for neighboor dependent substitution processes.
Mathematical Biosciences, 224(2):101–108, 2010.

Références II

T.H. Jukes and C.R. Cantor.
Mammalian protein metabolism, chapter Evolution of Protein Molecules, pages 21–132.
Academic Press, New York, 1969.

A. W. van der Vaart.
Asymptotic Statistics.
Cambridge University Press, 1998.