Rappels sur les modèles

Formation phylogénie – Institut Pasteur

Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558
Université Claude Bernard – Lyon 1

1er octobre 2015

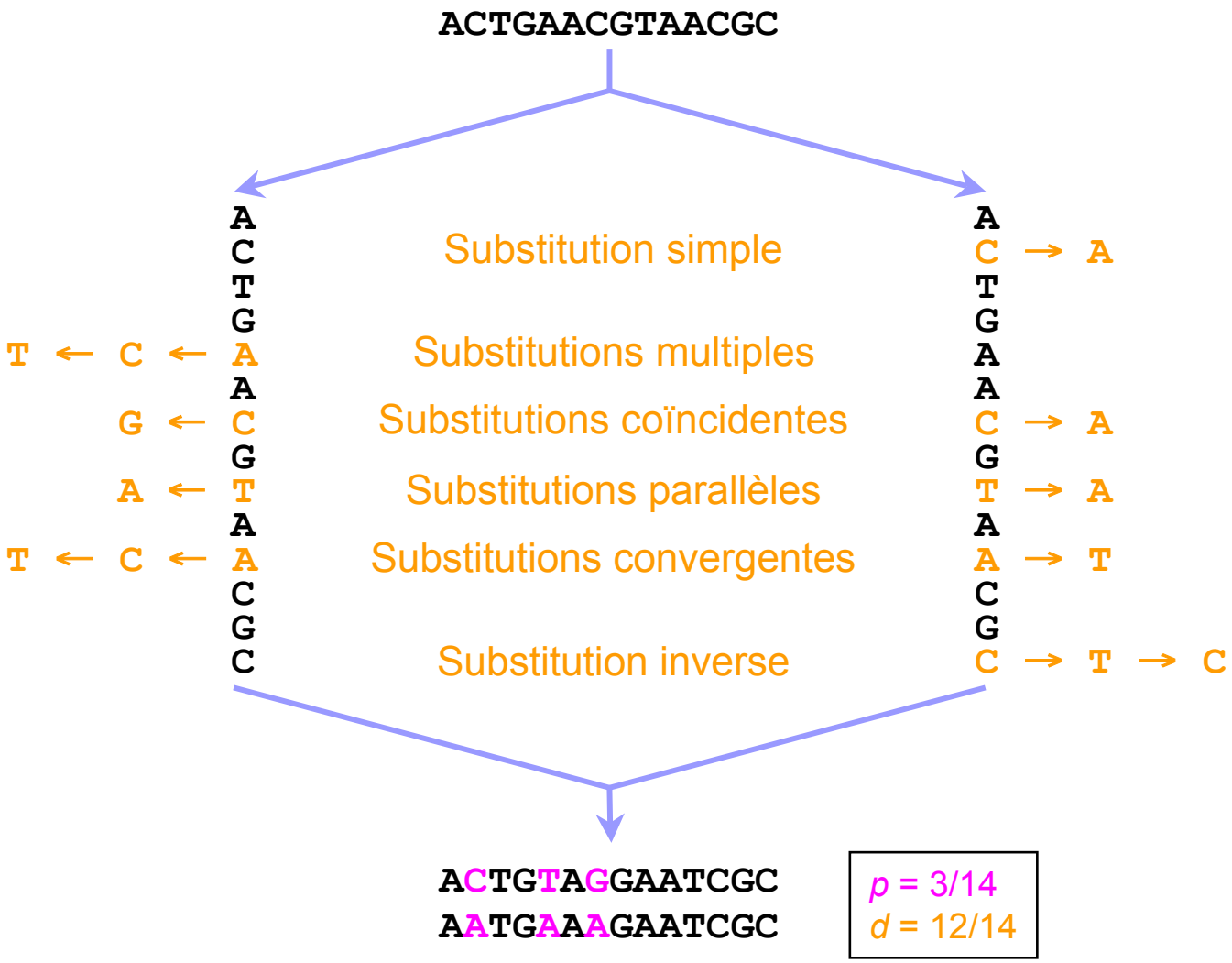
Divergence observée

- Appelée p (ou p-distance), c'est l'estimation la plus simple de la distance entre deux séquences :

$$p = n/\ell$$

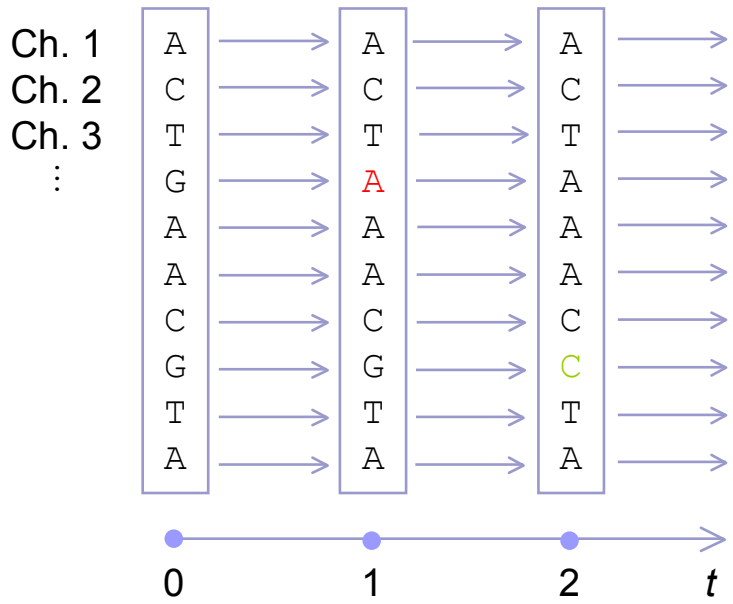
  avec n le nombre total de substitutions et ℓ le nombre de sites homologues comparés.
- La distance évolutive réelle (d) est généralement supérieure à la divergence observée (p).
- En faisant des hypothèses sur la nature du processus évolutif, il est possible d'estimer d à partir de p.

Types de substitutions



Modèlisation markovienne de l'évolution

- Modèles pour les séquences d'ADN ou de protéines.
- Hypothèses des modèles courants :
  - Temps continu.
  - Homogénéité.
  - Distribution stationnaire :
    - Stationnarité atteinte dès la racine.
  - Réversibilité.
  - Indépendance des sites.
  - Uniformité du processus :
    - Une seule matrice de transition.



Évolution des sites d'une séquence d'ADN selon un processus markovien

Taux instantanés

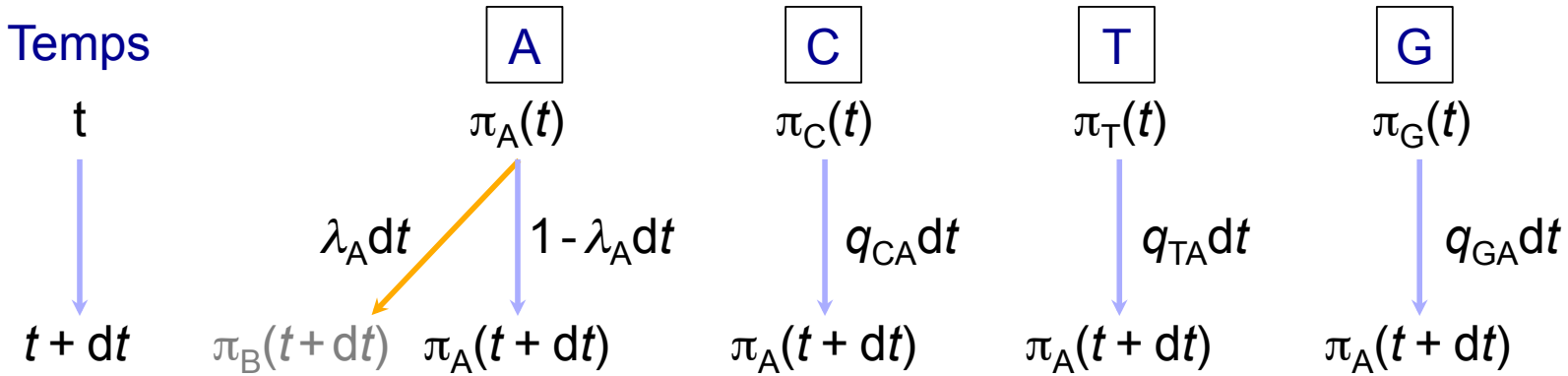
- Soit q_{ij} $(i \neq j)$ le *taux de substitution instantané* d'un nucléotide i vers un nucléotide j $(i, j \in \{\text{A, C, T, G}\})$.
- Dans ce cas le *taux de changement instantané* d'un nucléotide i est défini comme $\lambda_i = \sum_{j \neq i} q_{ij}$.
- L'ensemble des taux de substitutions et des taux de changements peuvent être regroupés dans une matrice $\mathbf{Q} = (q_{ij})$ telle que :

$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\text{A}} & q_{\text{AC}} & q_{\text{AT}} & q_{\text{AG}} \\ q_{\text{CA}} & -\lambda_{\text{C}} & q_{\text{CT}} & q_{\text{CG}} \\ q_{\text{TA}} & q_{\text{TC}} & -\lambda_{\text{T}} & q_{\text{TG}} \\ q_{\text{GA}} & q_{\text{GC}} & q_{\text{GT}} & -\lambda_{\text{G}} \end{pmatrix}$$

Les sommes en ligne de $\mathbf{Q}$ sont égales à 0.

Dynamique de fréquences des bases

- Soit $\boldsymbol{\pi}(t) = (\pi_{\mathrm{A}}(t), \pi_{\mathrm{C}}(t), \pi_{\mathrm{T}}(t), \pi_{\mathrm{G}}(t))$ le vecteur ligne des fréquences des nucléotides au temps t.
- Au temps $t + \mathrm{d}t$ la fréquence du nucléotide A peut se calculer de la façon suivante :



avec $\mathrm{B} = \{\mathrm{C}, \mathrm{T}, \mathrm{G}\}$.

Généralisation

- Les fréquences des quatre nucléotides A, C, T et G au temps $t + \mathrm{d}t$ sont données par le système d'équations différentielles :

$$\pi_{\mathrm{A}}(t+\mathrm{d}t) = \pi_{\mathrm{A}}(t)(1-\lambda_{\mathrm{A}}\mathrm{d}t) + \pi_{\mathrm{C}}(t)q_{\mathrm{CA}}\mathrm{d}t + \pi_{\mathrm{T}}(t)q_{\mathrm{TA}}\mathrm{d}t + \pi_{\mathrm{G}}(t)q_{\mathrm{GA}}\mathrm{d}t$$
$$\pi_{\mathrm{C}}(t+\mathrm{d}t) = \pi_{\mathrm{C}}(t)(1-\lambda_{\mathrm{C}}\mathrm{d}t) + \pi_{\mathrm{A}}(t)q_{\mathrm{AC}}\mathrm{d}t + \pi_{\mathrm{T}}(t)q_{\mathrm{TC}}\mathrm{d}t + \pi_{\mathrm{G}}(t)q_{\mathrm{GC}}\mathrm{d}t$$
$$\pi_{\mathrm{T}}(t+\mathrm{d}t) = \pi_{\mathrm{T}}(t)(1-\lambda_{\mathrm{T}}\mathrm{d}t) + \pi_{\mathrm{A}}(t)q_{\mathrm{AT}}\mathrm{d}t + \pi_{\mathrm{C}}(t)q_{\mathrm{CT}}\mathrm{d}t + \pi_{\mathrm{G}}(t)q_{\mathrm{GT}}\mathrm{d}t$$
$$\pi_{\mathrm{G}}(t+\mathrm{d}t) = \pi_{\mathrm{G}}(t)(1-\lambda_{\mathrm{G}}\mathrm{d}t) + \pi_{\mathrm{A}}(t)q_{\mathrm{AG}}\mathrm{d}t + \pi_{\mathrm{C}}(t)q_{\mathrm{CG}}\mathrm{d}t + \pi_{\mathrm{T}}(t)q_{\mathrm{TG}}\mathrm{d}t$$

- Soit, sous forme matricielle :

$$\boldsymbol{\pi}(t+\mathrm{d}t) = \boldsymbol{\pi}(t) + \boldsymbol{\pi}(t)\mathbf{Q}\mathrm{d}t \Leftrightarrow$$
$$\frac{\mathrm{d}\boldsymbol{\pi}(t)}{\mathrm{d}t} = \boldsymbol{\pi}(t)\mathbf{Q}$$

Probabilités de transition

- La résolution de l'équation précédente donne :

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\mathrm{e}^{\mathbf{Q}t}$$

où $\boldsymbol{\pi}(0) = (\pi_{\mathrm{A}}(0), \pi_{\mathrm{C}}(0), \pi_{\mathrm{T}}(0), \pi_{\mathrm{G}}(0))$ est le vecteur ligne des fréquences ancestrales des nucléotides.

- On pose $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$ la matrice des *probabilités de transition* du processus de Markov, telle que :

$$\mathbf{P}(t) = \begin{pmatrix} p_{\mathrm{AA}}(t) & p_{\mathrm{AC}}(t) & p_{\mathrm{AT}}(t) & p_{\mathrm{AG}}(t) \\ p_{\mathrm{CA}}(t) & p_{\mathrm{CC}}(t) & p_{\mathrm{CT}}(t) & p_{\mathrm{CG}}(t) \\ p_{\mathrm{TA}}(t) & p_{\mathrm{TC}}(t) & p_{\mathrm{TT}}(t) & p_{\mathrm{TG}}(t) \\ p_{\mathrm{GA}}(t) & p_{\mathrm{GC}}(t) & p_{\mathrm{GT}}(t) & p_{\mathrm{GG}}(t) \end{pmatrix}$$

Les sommes en ligne de $\mathbf{P}(t)$ sont égales à 1.

Stationnarité

- La *distribution stationnaire* $\boldsymbol{\pi} = (\pi_i)$ correspond à la distribution vers laquelle un processus de Markov converge lorsque $t \to \infty$:

$$\lim_{t \to \infty} \pi_i(t) = \pi_i$$

  Dans le cas des séquences nucléotidiques, les valeurs de π_i sont appelées *fréquences des bases à l'équilibre.*

- L'existence d'une distribution stationnaire implique que :

$$\boldsymbol{\pi}\mathbf{P}(t) = \boldsymbol{\pi}, \ \forall t \geq 0$$

  ou son équivalent :

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$$

Réversibilité

- Un processus de Markov est dit *réversible* si, lorsque la stationnarité est atteinte, on a :

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \ \forall i, j \in \{\text{A, C, T, G}\}$$

  À l'équilibre, la quantité de changement $i \rightarrow j$ est égale à la quantité de changement $j \rightarrow i$.
- Sous l'hypothèse de réversibilité, il est possible d'écrire l'expression des valeurs de q_{ij} comme :

$$q_{ij} = \pi_j s_{ij} \quad (i \neq j)$$

  avec $s_{ij} = s_{ji}$ un terme symétrique, appelé paramètre *d'échangeabilité* entre i et j.

Matrices S et $\mathbf{\Pi}$

- Sous l'hypothèse de réversibilité, l'expression de $\mathbf{Q}$ peut s'écrire comme étant le produit :

$$\mathbf{Q} = \mathbf{S\Pi} = \begin{pmatrix} \cdot & \alpha & \beta & \gamma \\ \alpha & \cdot & \delta & \epsilon \\ \beta & \delta & \cdot & \eta \\ \gamma & \epsilon & \eta & \cdot \end{pmatrix} \times \begin{pmatrix} \pi_{\mathrm{A}} & 0 & 0 & 0 \\ 0 & \pi_{\mathrm{C}} & 0 & 0 \\ 0 & 0 & \pi_{\mathrm{T}} & 0 \\ 0 & 0 & 0 & \pi_{\mathrm{G}} \end{pmatrix}$$

avec $\mathbf{S}$ la matrice des écheangeabilités entre nucléotides et $\mathbf{\Pi} = \mathrm{diag}(\pi_i)$ la matrice diagonale contenant les valeurs des fréquences des bases à l'équilibre.

Expression de Q

- Au moyen du produit matriciel précédent, on en déduit l'expression de $\mathbf{Q}$:

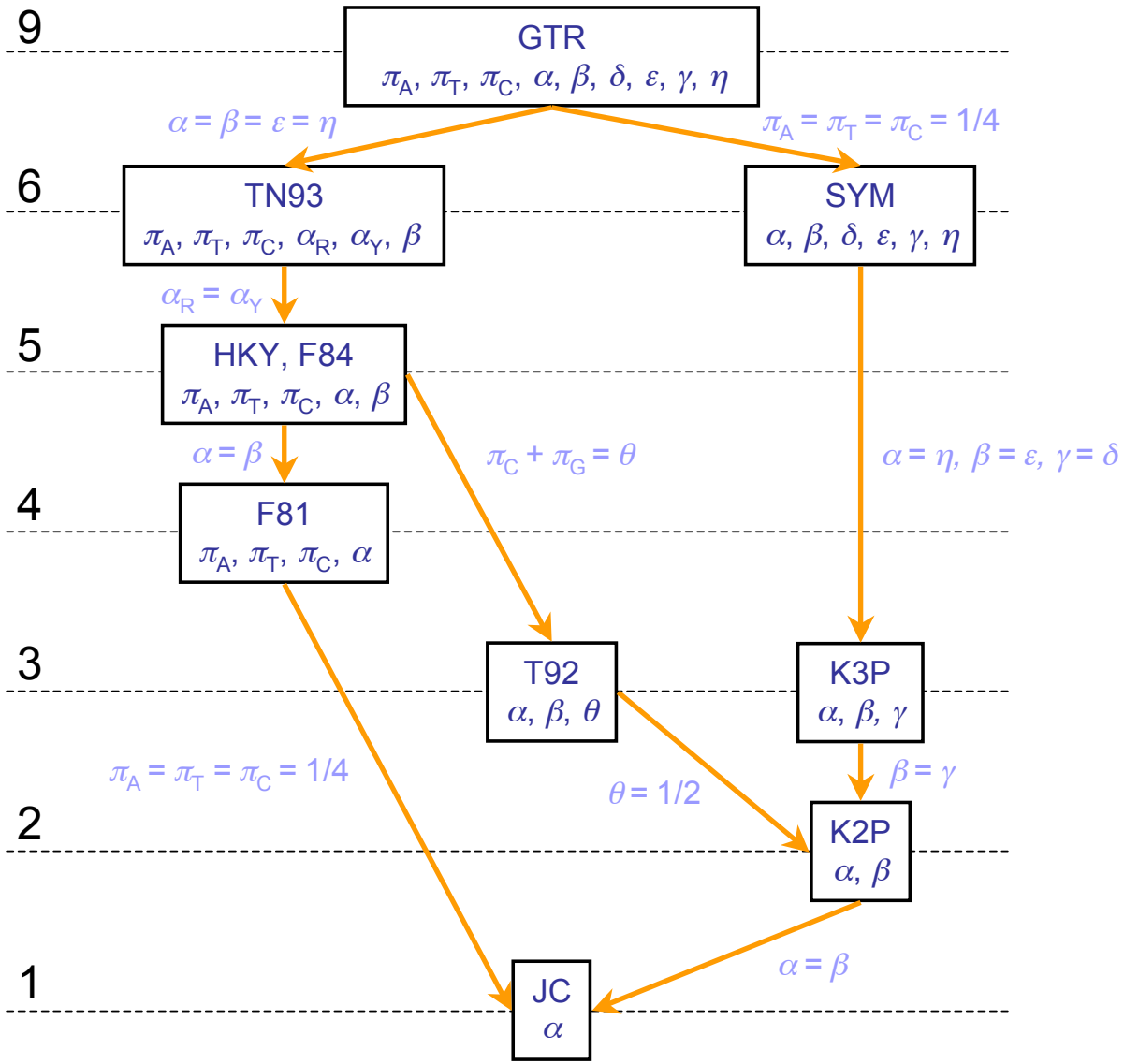
$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{C}}\alpha & \pi_{\mathrm{T}}\beta & \pi_{\mathrm{G}}\gamma \\ \pi_{\mathrm{A}}\alpha & -\lambda_{\mathrm{C}} & \pi_{\mathrm{T}}\delta & \pi_{\mathrm{G}}\epsilon \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\delta & -\lambda_{\mathrm{T}} & \pi_{\mathrm{G}}\eta \\ \pi_{\mathrm{A}}\gamma & \pi_{\mathrm{C}}\epsilon & \pi_{\mathrm{T}}\eta & -\lambda_{\mathrm{G}} \end{pmatrix}$$

$$\text{avec} \begin{cases} \lambda_{\mathrm{A}} = \pi_{\mathrm{C}}\alpha + \pi_{\mathrm{T}}\beta + \pi_{\mathrm{G}}\gamma \\ \lambda_{\mathrm{C}} = \pi_{\mathrm{A}}\alpha + \pi_{\mathrm{T}}\delta + \pi_{\mathrm{G}}\epsilon \\ \lambda_{\mathrm{T}} = \pi_{\mathrm{A}}\beta + \pi_{\mathrm{C}}\delta + \pi_{\mathrm{G}}\eta \\ \lambda_{\mathrm{G}} = \pi_{\mathrm{A}}\gamma + \pi_{\mathrm{C}}\epsilon + \pi_{\mathrm{T}}\eta \end{cases}$$

Soit neuf paramètres à estimer :

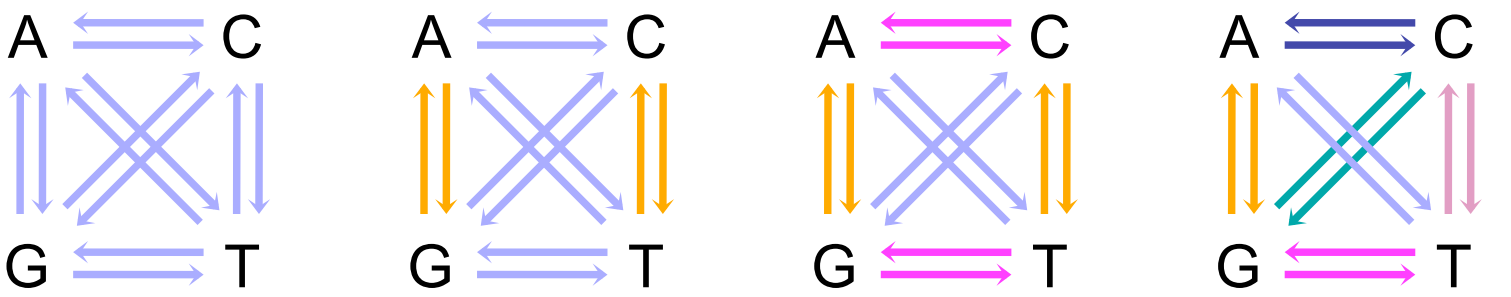
- Modèle GTR (*Generalised Time Reversible*) ou REV.

Imbrication des modèles



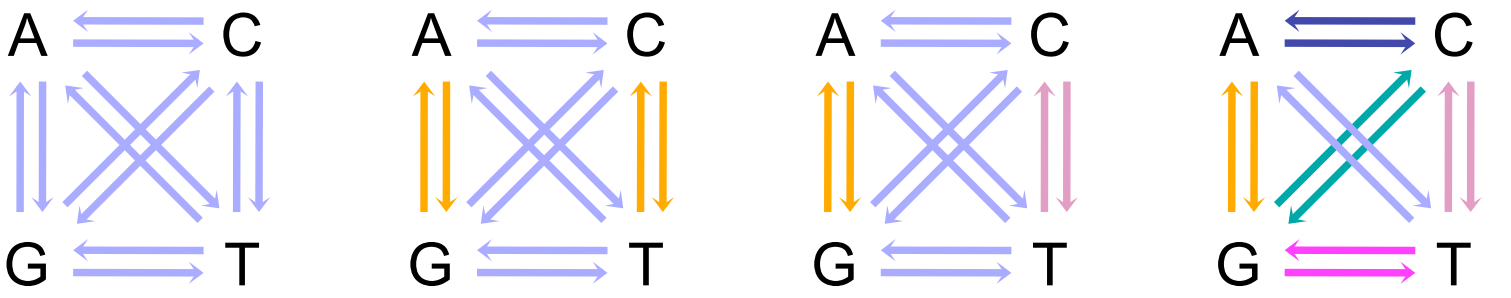
Paramètres des modèles

$\pi_A = \pi_C = \pi_T = \pi_G = 1/4$



JC	K2P	K3P	SYM
(α)	(α, β)	(α, β, γ)	(α, β, δ, ε, γ, η)

$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$



F81	HKY, F84	TN93	GTR
(α)	(α, β)	(α_R, α_Y, β)	(α, β, δ, ε, γ, η)

Calcul de la distance évolutive

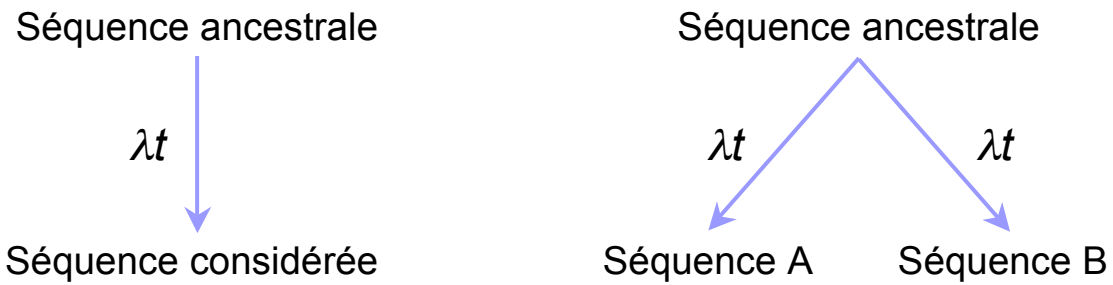
- Soit λ, le *taux global de substitions* dans une séquence. Sous l'hypothèse de réversibilité, ce taux est égal à :

$$\lambda = \sum_i \pi_i \lambda_i, \;\; i \in \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$$

avec λ_i le taux de changement instantané d'un nucléotide en n'importe lequel des trois autres.

- Dans ce cas, la distance évolutive entre deux séquences est donnée par la formule :

$$d = 2\lambda t = 2 \sum_i \pi_i \lambda_i t$$



Quelques distances I

- Modèle de Jukes et Cantor (1969) – JC :

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

- Modèle de Kimura (1980) – K2P :

$$d = -\frac{1}{2}\ln(1 - 2r - v) - \frac{1}{4}\ln(1 - 2v)$$

avec r la fréquence des transitions et v la fréquence des transversions observées entre les deux séquences ($p = r + v$).

Quelques distances II

- Modèle de Felsenstein (1981) – F81 :

$$d = -a \ln\left(1 - \frac{p}{a}\right)$$

avec $a = 1 - \pi_{\mathrm{A}}^2 - \pi_{\mathrm{C}}^2 - \pi_{\mathrm{T}}^2 - \pi_{\mathrm{G}}^2$.

- Modèle de Felsenstein (1984) – F84 :

$$d = -2a_1 \ln\left[1 - \frac{r}{2a_1} - \frac{(a_1 - a_2)v}{2a_1 a_3}\right]$$

$$\text{avec } \begin{cases} a_1 = \dfrac{\pi_{\mathrm{A}}\pi_{\mathrm{G}}}{\pi_{\mathrm{A}} + \pi_{\mathrm{G}}} + \dfrac{\pi_{\mathrm{C}}\pi_{\mathrm{T}}}{\pi_{\mathrm{C}} + \pi_{\mathrm{T}}} \\ a_2 = \pi_{\mathrm{A}}\pi_{\mathrm{G}} + \pi_{\mathrm{C}}\pi_{\mathrm{T}} \\ a_3 = (\pi_{\mathrm{A}} + \pi_{\mathrm{G}})(\pi_{\mathrm{C}} + \pi_{\mathrm{T}}) \end{cases}$$

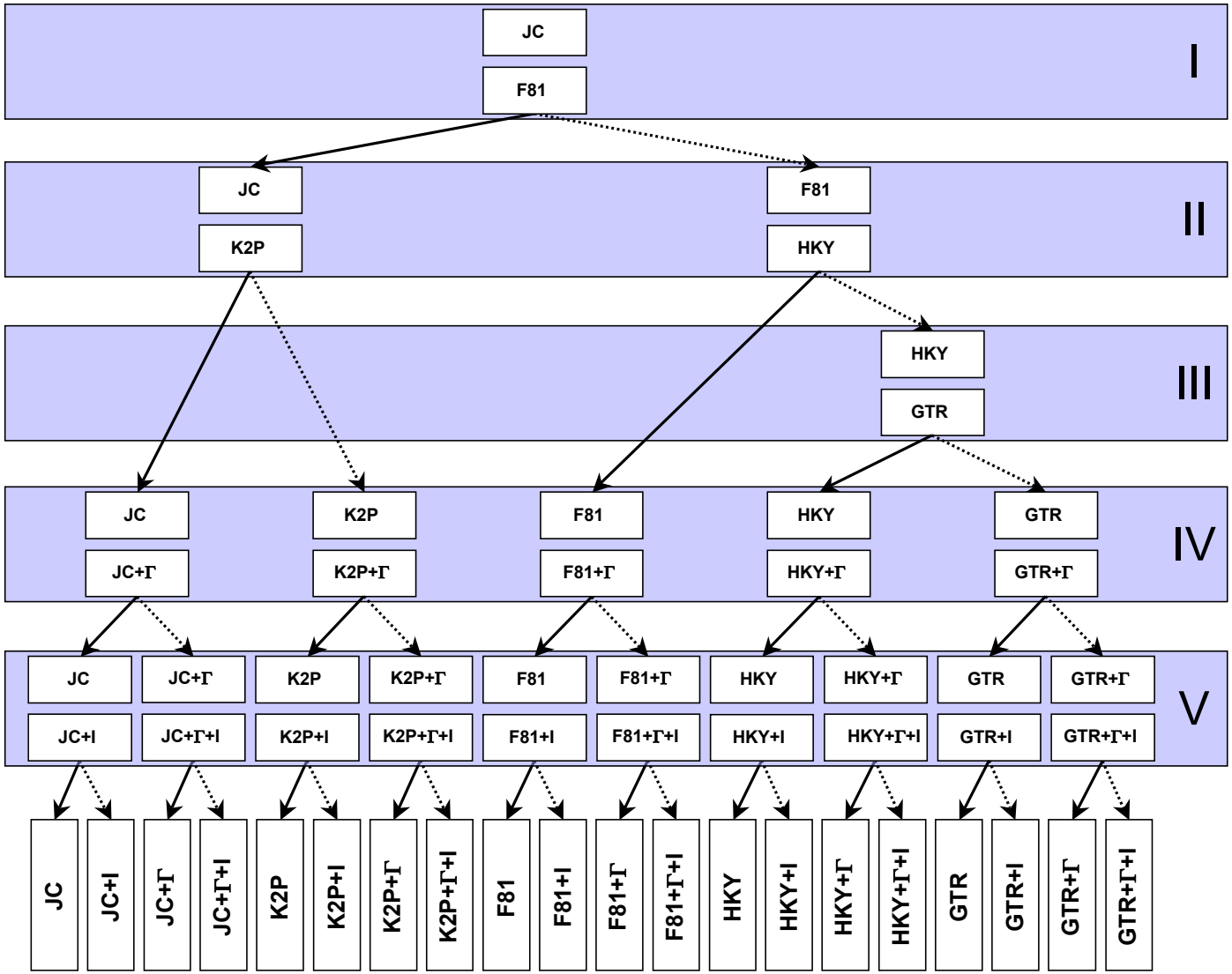
Likelihood Ratio Test (LRT)

- Soient M_0 et M_1 deux modèles caractérisés par leurs vecteurs de paramètres $\boldsymbol{\vartheta}_0$ et $\boldsymbol{\vartheta}_1$ tels que $k_0 = \dim(\boldsymbol{\vartheta}_0)$ et $k_1 = \dim(\boldsymbol{\vartheta}_1)$:
  - M_0 doit être *imbriqué* dans M_1 ($k_0 < k_1$).
- Le rapport des vraisemblances est donné par :

$$\Lambda = 2 \ln \left[\frac{L(\boldsymbol{\vartheta}_1)}{L(\boldsymbol{\vartheta}_0)} \right] = 2[\ln L(\boldsymbol{\vartheta}_1) - \ln L(\boldsymbol{\vartheta}_0)]$$

  avec $L(\boldsymbol{\vartheta}_0)$ et $L(\boldsymbol{\vartheta}_1)$ les vraisemblances associés à M_0 et M_1.
- Pour le calcul du test proprement dit, on considère que $\Lambda \sim \chi^2(k_1 - k_0)$.

Arbre de décision du LRT



Akaike Information Criterion (AIC)

- Test AIC standard :

$$\mathrm{AIC} = -2 \ln L(\boldsymbol{\vartheta}) + 2k$$

avec $k = \dim(\boldsymbol{\vartheta})$ le nombre de paramètres du modèle.

- Test AICc, incluant une correction par la taille de l'échantillon :

$$\mathrm{AICc} = -2 \ln L(\boldsymbol{\vartheta}) + \frac{2k(k+1)}{\ell - k - 1}$$

avec ℓ la longueur de l'alignement.

- Dans les deux cas, sélection du modèle présentant la plus faible valeur au test.

Bayesian Information Criterion (BIC)

- Test BIC standard :

$$\text{BIC} = -2 \ln L(\boldsymbol{\vartheta}) + k \ln \ell$$

- Comme dans le cas de l'AIC, sélection du modèle présentant la plus faible valeur au test.
- Approximation du test de comparaison de modèles utilisant les Facteurs de Bayes :

$$2 \ln \text{BF}_{10} \approx \text{BIC}_1 - \text{BIC}_0$$