Modèles d'évolution

Analyse de séquences génomiques et phylogénie

Guy Perrière

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

3-6 avril 2017

Divergence observée

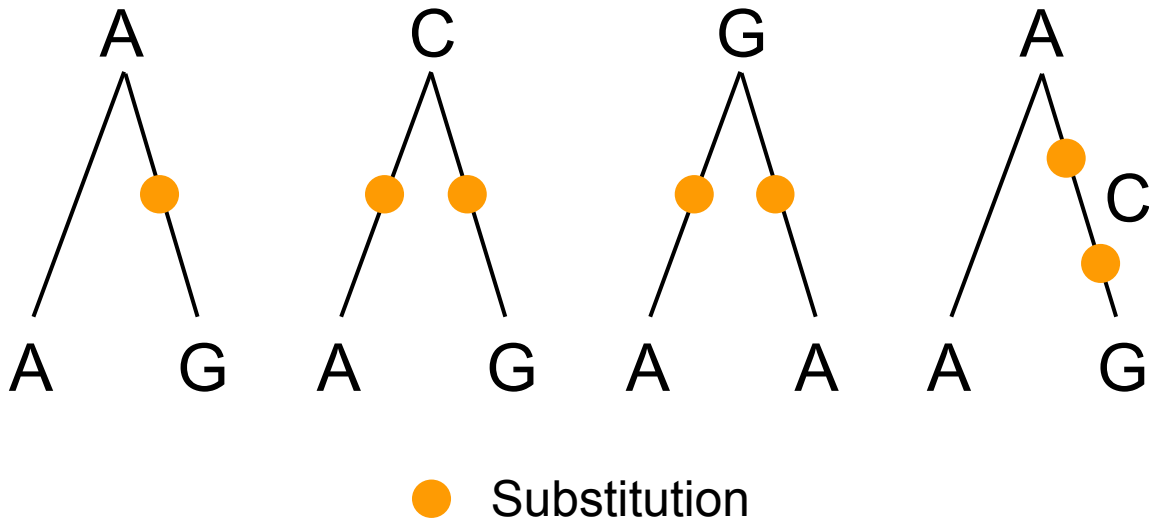
- Appelée p (ou p-distance), c'est l'estimation la plus simple de la distance entre deux séquences :

$$p = n/\ell$$

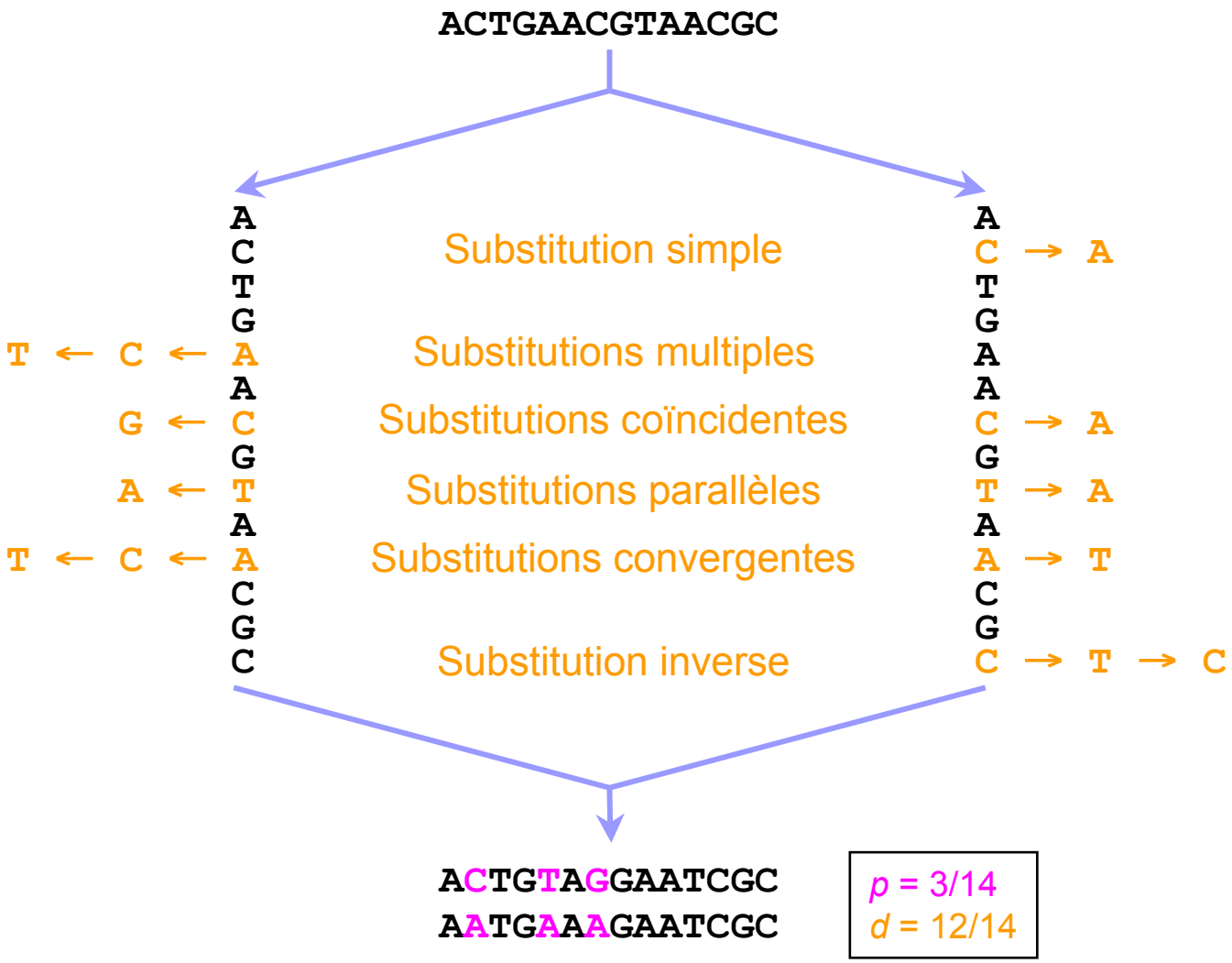
  avec n le nombre total de substitutions et ℓ le nombre de sites homologues comparés.
- Variation pour deux séquences de composition homogène :
  - Pour l'ADN : $0 \leq p \leq 0.75$.
  - Pour les protéines : $0 \leq p \leq 0.95$.

Substitutions multiples

- 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.

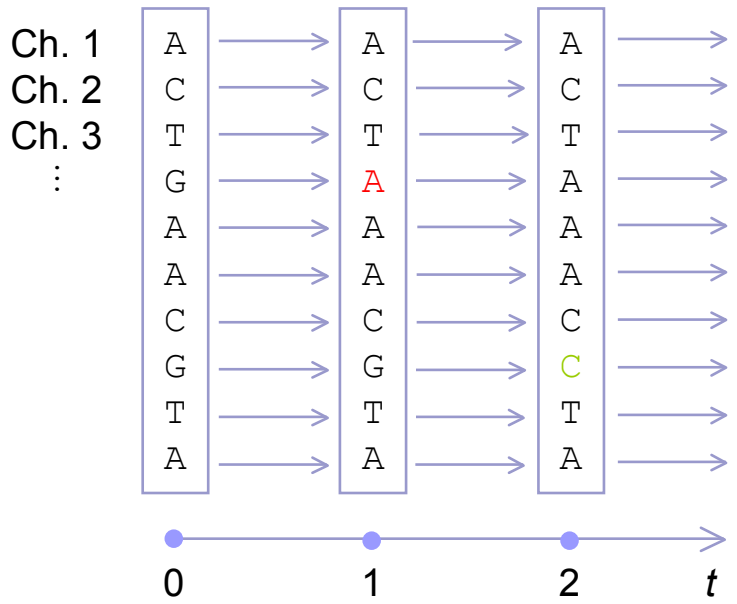


Types de substitutions



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

Nombre de substitutions

- On pose $\Omega = \{\text{A, C, T, G}\}$ l'ensemble des états possibles.
- Soit $\mathbf{N} = (n_{ij})$ $(i, j \in \Omega)$, la matrice contenant le nombre de substitutions $(i \neq j)$ et de conservations $(i = j)$ observées entre deux séquences alignées :

$$\mathbf{N} = \begin{pmatrix} n_{\text{AA}} & n_{\text{AC}} & n_{\text{AT}} & n_{\text{AG}} \\ n_{\text{CA}} & n_{\text{CC}} & n_{\text{CT}} & n_{\text{CG}} \\ n_{\text{TA}} & n_{\text{TC}} & n_{\text{TT}} & n_{\text{TG}} \\ n_{\text{GA}} & n_{\text{GC}} & n_{\text{GT}} & n_{\text{GG}} \end{pmatrix}$$

- Le nombre total de substitutions observées n est tel que :

$$n = \sum_{i \neq j} n_{ij}$$

Fréquence des substitutions

- Soit $\mathbf{F} = (f_{ij})$ $(i, j \in \Omega)$, la matrice contenant les fréquences des substitutions $(i \neq j)$ et des conservations $(i = j)$ observées entre deux séquences alignées :

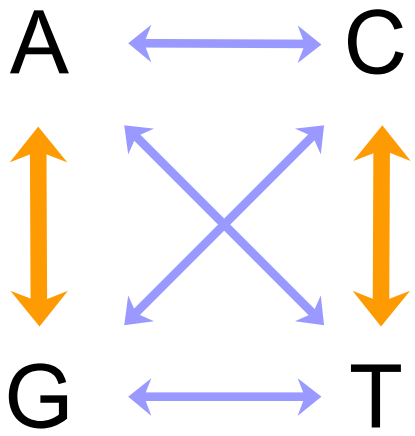
$$\mathbf{F} = \begin{pmatrix} f_{\mathrm{AA}} & f_{\mathrm{AC}} & f_{\mathrm{AT}} & f_{\mathrm{AG}} \\ f_{\mathrm{CA}} & f_{\mathrm{CC}} & f_{\mathrm{CT}} & f_{\mathrm{CG}} \\ f_{\mathrm{TA}} & f_{\mathrm{TC}} & f_{\mathrm{TT}} & f_{\mathrm{TG}} \\ f_{\mathrm{GA}} & f_{\mathrm{GC}} & f_{\mathrm{GT}} & f_{\mathrm{GG}} \end{pmatrix}$$

- Soit ℓ le nombre de sites homologues comparés, dans ce cas :

$$f_{ij} = \frac{n_{ij}}{\ell} \quad \text{et} \quad p = \sum_{i \neq j} f_{ij} = \frac{n}{\ell}$$

Transitions et transversions

- Beaucoup de modèles font la distinction entre les substitutions de type transitions et celles de type transversions :



- Soit r la fréquence des transitions et v celle des transversions, telles que :

$$r = r_{\mathrm{R}} + r_{\mathrm{Y}} = f_{\mathrm{AG}} + f_{\mathrm{GA}} + f_{\mathrm{CT}} + f_{\mathrm{TC}}$$
$$v = f_{\mathrm{AC}} + f_{\mathrm{CA}} + f_{\mathrm{AT}} + f_{\mathrm{TA}} + f_{\mathrm{CG}} + f_{\mathrm{GC}} + f_{\mathrm{GT}} + f_{\mathrm{TG}}$$

Taux instantanés

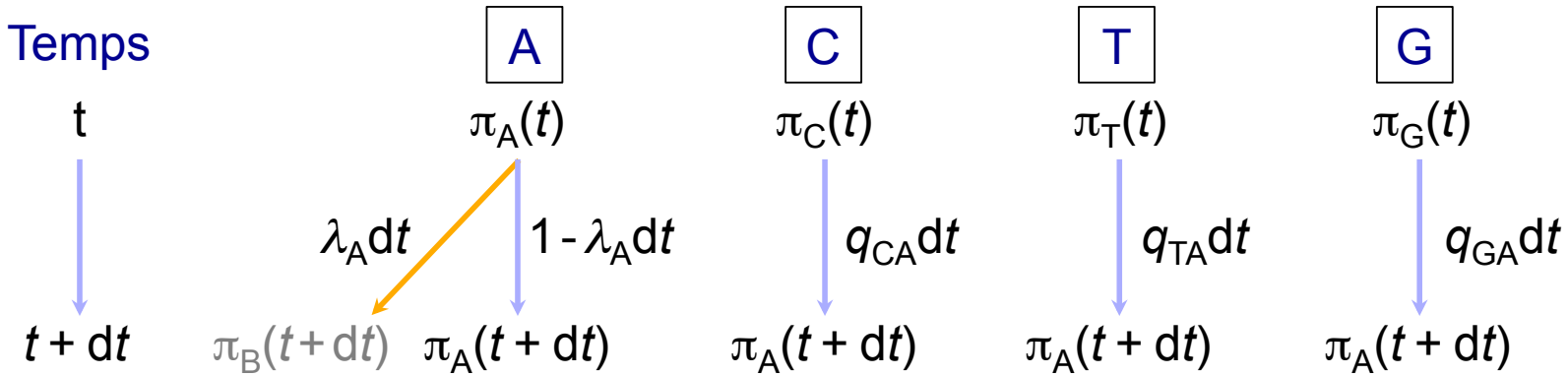
- Soit q_{ij} $(i \neq j)$ le *taux de substitution instantané* d'un nucléotide i vers un nucléotide j $(i, j \in \Omega)$.
- Dans ce cas le *taux de changement instantané* d'un nucléotide i est défini comme $\lambda_i = \sum_{j \neq i} q_{ij}$.
- L'ensemble des taux de substitutions et des taux de changements peuvent être regroupés dans une matrice $\mathbf{Q} = (q_{ij})$ telle que :

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

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 :

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

- Soit, sous forme matricielle :

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

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 :

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \ \forall i, j \in \Omega$$

  À l'équilibre, la quantité de changement $i \to j$ est égale à la quantité de changement $j \to i$.
- 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 échangeabilités 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 $\mathbf{Q}$:

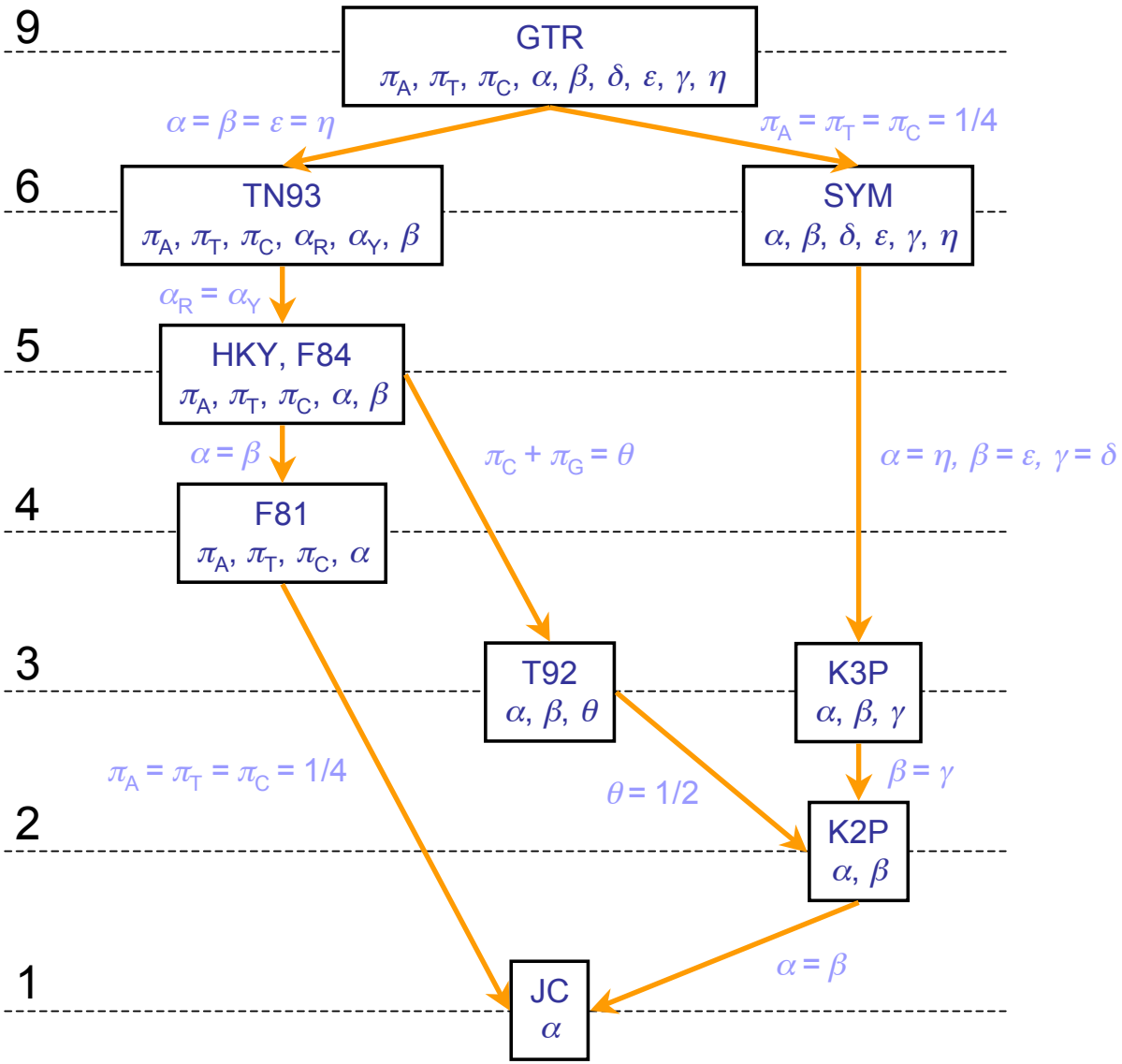
$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{C}}\alpha & \pi_{\mathrm{T}}\beta & \pi_{\mathrm{G}}\gamma \\ \pi_{\mathrm{A}}\alpha & -\lambda_{\mathrm{C}} & \pi_{\mathrm{T}}\delta & \pi_{\mathrm{G}}\epsilon \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\delta & -\lambda_{\mathrm{T}}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

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

Soit neuf paramètres à estimer :

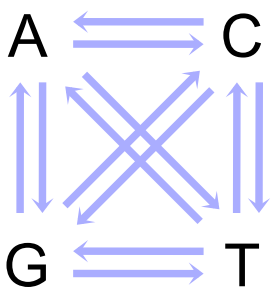
- Modèle GTR (*Generalised Time Reversible*) ou REV.

Imbrication des modèles

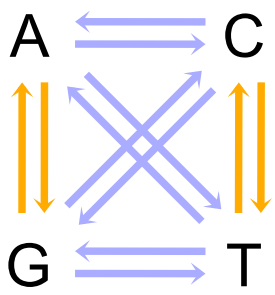


Paramètres des modèles

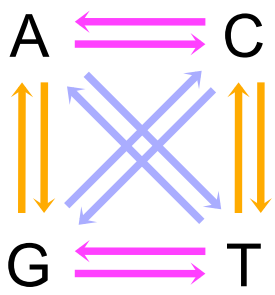
$\pi_A = \pi_C = \pi_T = \pi_G = 1/4$



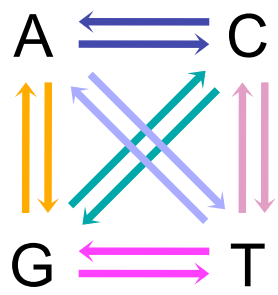
JC
(α)



K2P
(α, β)

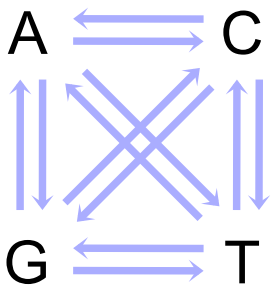


K3P
(α, β, γ)

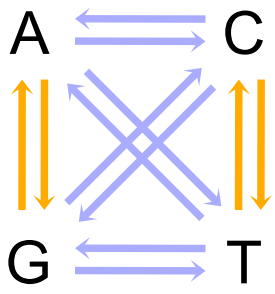


SYM
(α, β, δ, ε, γ, η)

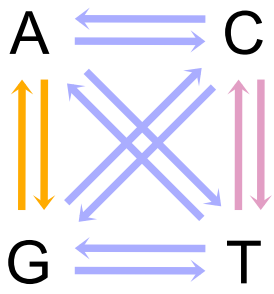
$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$



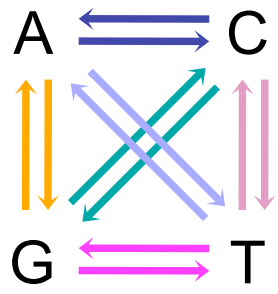
F81
(α)



HKY, F84
(α, β)



TN93
(α_R, α_Y, β)



GTR
(α, β, δ, ε, γ, η)

Calcul de la distance évolutive

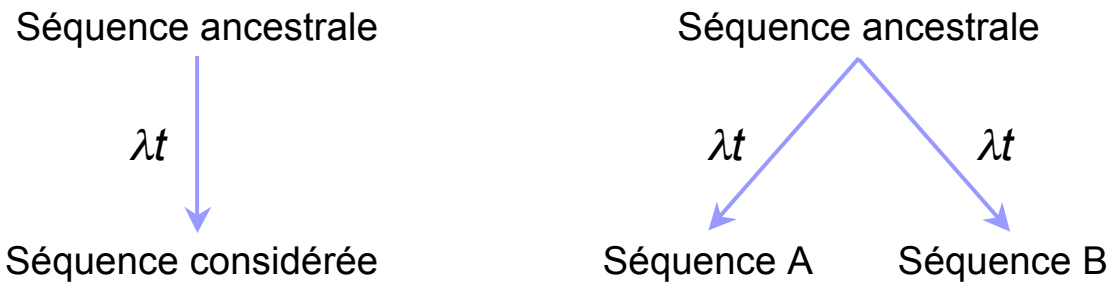
- Soit λ, le *taux global de substitions* dans une séquence. Sous l'hypothèse de réversibilité, ce taux est égal à :

$$\lambda = \sum_i \pi_i \lambda_i, \;\; i \in \Omega$$

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 :

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



Normalisation

- Par convention, les valeurs des taux instantanés sont normalisées de façon à ce que :

$$\lambda = \sum_i \pi_i \lambda_i = 1$$

- Sous cette contrainte, la distance évolutive entre deux séquences est assimilable au temps écoulé :

$$d = 2\lambda t = 2t$$

Modèle de Jukes et Cantor

- Une seule échangeabilité (α), identique pour chacun des quatre nucléotides.
- Fréquences à l'équilibre : $\pi_{\mathrm{A}} = \pi_{\mathrm{C}} = \pi_{\mathrm{T}} = \pi_{\mathrm{G}} = 1/4$.
- Matrices $\mathbf{Q}$ et $\mathbf{P}(t)$:

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \alpha & \alpha & \alpha \\ \alpha & -\lambda & \alpha & \alpha \\ \alpha & \alpha & -\lambda & \alpha \\ \alpha & \alpha & \alpha & -\lambda \end{pmatrix} \quad \mathbf{P}(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

avec $p_0(t) + 3p_1(t) = 1$.

- Taux global de substitutions : $\lambda = \sum_i \pi_i \lambda_i = 3\alpha$.

Résolution

- Le calcul de $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$ permet de déterminer que :

$$p_0(t) = \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4\alpha t} \quad \text{et} \quad p_1(t) = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\alpha t}$$

- Élimination de t et introduction de d en utilisant la relation $d = 2\lambda t = 6\alpha t$.
- Introduction de la divergence observée entre deux séquences $p = 3p_1(2t)$.
- Formule de Jukes et Cantor pour le calcul de la distance évolutive :

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

Autres distances I

- Modèle de Kimura à deux paramètres(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 ($p = r + v$).

- Modèle de Felsenstein (1981) – F81 :

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

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

Autres distances II

- Modèle de Felsenstein (1984) – F84 :

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

$$\text{avec} \begin{cases} a_1 = \dfrac{\pi_A \pi_G}{\pi_A + \pi_G} + \dfrac{\pi_C \pi_T}{\pi_C + \pi_T} \\ a_2 = \pi_A \pi_G + \pi_C \pi_T \\ a_3 = (\pi_A + \pi_G)(\pi_C + \pi_T) \end{cases}$$

Autres distances III

- Modèle de Tamura et Nei (1993) – TN93 :

$$d = \frac{2\pi_{\mathrm{T}}\pi_{\mathrm{C}}}{\pi_{\mathrm{Y}}}(a_1 - \pi_{\mathrm{R}} b) + \frac{2\pi_{\mathrm{A}}\pi_{\mathrm{G}}}{\pi_{\mathrm{R}}}(a_2 - \pi_{\mathrm{Y}} b) + 2\pi_{\mathrm{Y}}\pi_{\mathrm{R}} b$$

$$\text{avec} \begin{cases} a_1 = -\ln\left(1 - \dfrac{\pi_{\mathrm{Y}}}{2\pi_{\mathrm{T}}\pi_{\mathrm{C}}} r_{\mathrm{Y}} - \dfrac{1}{2\pi_{\mathrm{Y}}} v\right) \\ a_2 = -\ln\left(1 - \dfrac{\pi_{\mathrm{R}}}{2\pi_{\mathrm{A}}\pi_{\mathrm{G}}} r_{\mathrm{R}} - \dfrac{1}{2\pi_{\mathrm{R}}} v\right) \\ b = -\ln\left(1 - \dfrac{1}{2\pi_{\mathrm{R}}\pi_{\mathrm{Y}}} v\right) \end{cases}$$

Modèle GTR

- Pas de solution analytique au calcul de $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$.
- La fréquence des substitutions $i \to j$ observées entre deux séquences au temps t est donnée par :

$$f_{ij}(t) = \pi_i p_{ij}(2t)$$

- Estimation des valeurs de π_i à partir des fréquences des bases dans les deux séquences considérées.
- Estimation des valeurs de $f_{ij}(t)$ en utilisant celles de celles de $\mathbf{F} = (f_{ij})$ (Diapo. 7).
- Calcul des valeurs de $p_{ij}(2t)$ à partir de l'équation précédente.

Utilité des modèles complexes

- Modélisent mieux l'évolution des séquences :
  - Plus proches de la réalité biologique.
- Séquences trop courtes :
  - Erreurs d'échantillonnage (valeurs de $d < 0$).
  - Variance importante.
- Séquences trop divergentes :
  - Méthodes à plus de quatre paramètres fréquemment inapplicables.
- Séquences peu divergentes :
  - Toutes les méthodes donnent des résultats comparables.

Le code génétique

I \ II	T	C	A	G	III
T	TTT Phe F	TCT Ser S	TAT Tyr Y	TGT Cys C	**T**
	TTC Phe F	TCC Ser S	TAC Tyr Y	TGC Cys C	**C**
	TTA Leu L	TCA Ser S	TAA Stop	TGA Stop	**A**
	TTG Leu L	TCG Ser S	TAG Stop	TGG Trp W	**G**
C	CTT Leu L	CCT Pro P	CAT His H	CGT Arg R	**T**
	CTC Leu L	CCC Pro P	CAC His H	CGC Arg R	**C**
	CTA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R	**A**
	CTG Leu L	CCG Pro P	CAG Gln Q	CGG Arg R	**G**
A	ATT Ile I	ACT Thr T	AAT Asn N	AGT Ser S	**T**
	ATC Ile I	ACC Thr T	AAC Asn N	AGC Ser S	**C**
	ATA Ile I	ACA Thr T	AAA Lys K	AGA Arg R	**A**
	ATG Met M	ACG Thr T	AAG Lys K	AGG Arg R	**G**
G	GTT Val V	GCT Ala A	GAT Asp D	GGT Gly G	**T**
	GTC Val V	GCC Ala A	GAC Asp D	GGC Gly G	**C**
	GTA Val V	GCA Ala A	GAA Glu E	GGA Gly G	**A**
	GTG Val V	GCG Ala A	GAG Glu E	GGG Gly G	**G**

Substitutions synonymes et non synonymes

- Exemple d'une transversion C $\rightarrow$ A dans le codon CTC :
  - En position I :

    $$\mathrm{CTC}^{\mathrm{Leu}} \rightarrow \mathrm{ATC}^{\mathrm{Ile}}$$

    soit une substitution *non synonyme* (ou *non silencieuse*).
  - En position III :

    $$\mathrm{CTC}^{\mathrm{Leu}} \rightarrow \mathrm{CTA}^{\mathrm{Leu}}$$

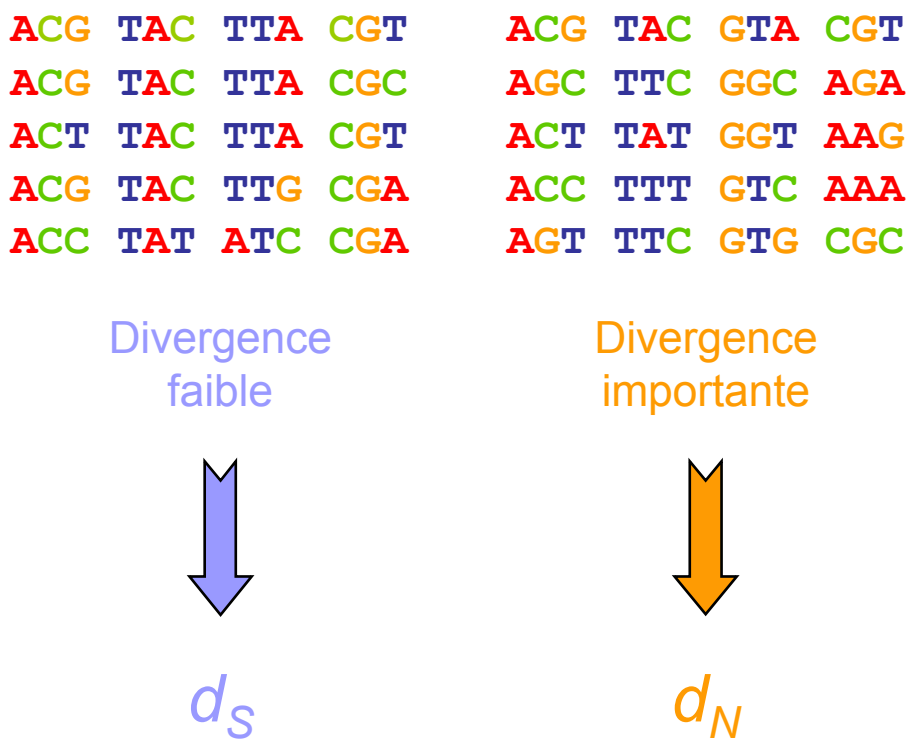
    soit une substitution *synonyme* (ou *silencieuse*).
- Toutes les substitutions touchant la position II des codons sont non synonymes.

Distances d_N et d_S

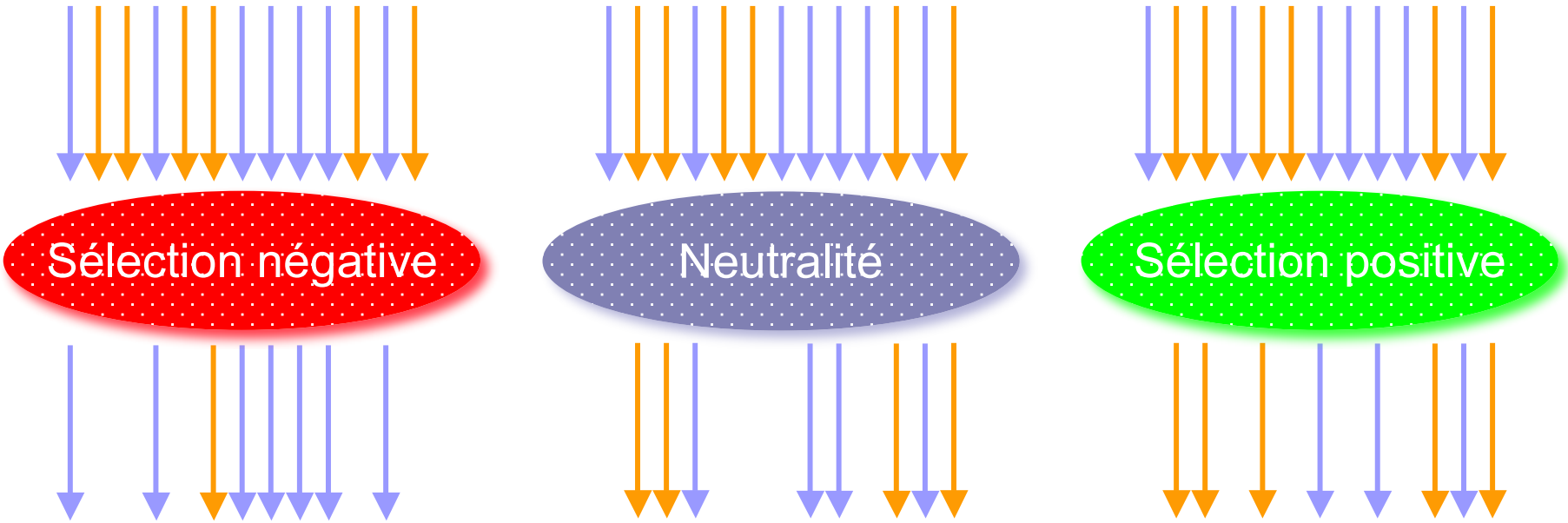
- Dans les gènes protéiques, il existe deux classes de sites ayant des vitesses évolutives différentes :
  - Substitutions non synonymes lentes.
  - Substitutions synonymes rapides.
  - L'hypothèse faite par les modèles d'évolution « classiques » que chaque site évolue en suivant le même processus est fausse.
- Calcul de deux distances évolutives différentes :
  - Distance non synonyme (d_N) :
    - Calcul à partir de p_N = nb. de substitutions non synonymes / nb. de sites non synonymes.
  - Distance synonyme (d_S) :
    - Calcul à partir de p_S = nb. de substitutions synonymes / nb. de sites synonymes.

Utilisation

- On se trouve fréquemment dans l'une ou l'autre de ces deux situations :
  - Séquences évolutivement peu distantes :
    - d_S est informatif, d_N ne l'est pas.
  - Séquences évolutivement très distantes :
    - d_S est saturé, d_N est informatif.



Sélection et neutralité



$d_N/d_S < 1$ $d_N/d_S \simeq 1$ $d_N/d_S > 1$

Substitutions synonymes
Substitutions non synonymes

Méthodes de comptage

- Gojobori et Nei (1986) :
  - Simplification des premières méthodes proposées par Miyata et Yasunaga (1980) et Perler *et al.* (1980).
  - Utilisation du modèle de Jukes et Cantor.
- Ina (1995) :
  - Amélioration de la méthode de Gojobori et Nei en distinguant les transitions des transversions.
  - Utilisation du modèle de Kimura à deux paramètres.
- Li, Wu et Luo (1985) :
  - Utilise la dégénérescence du code génétique et le modèle de Kimura à deux paramètres.

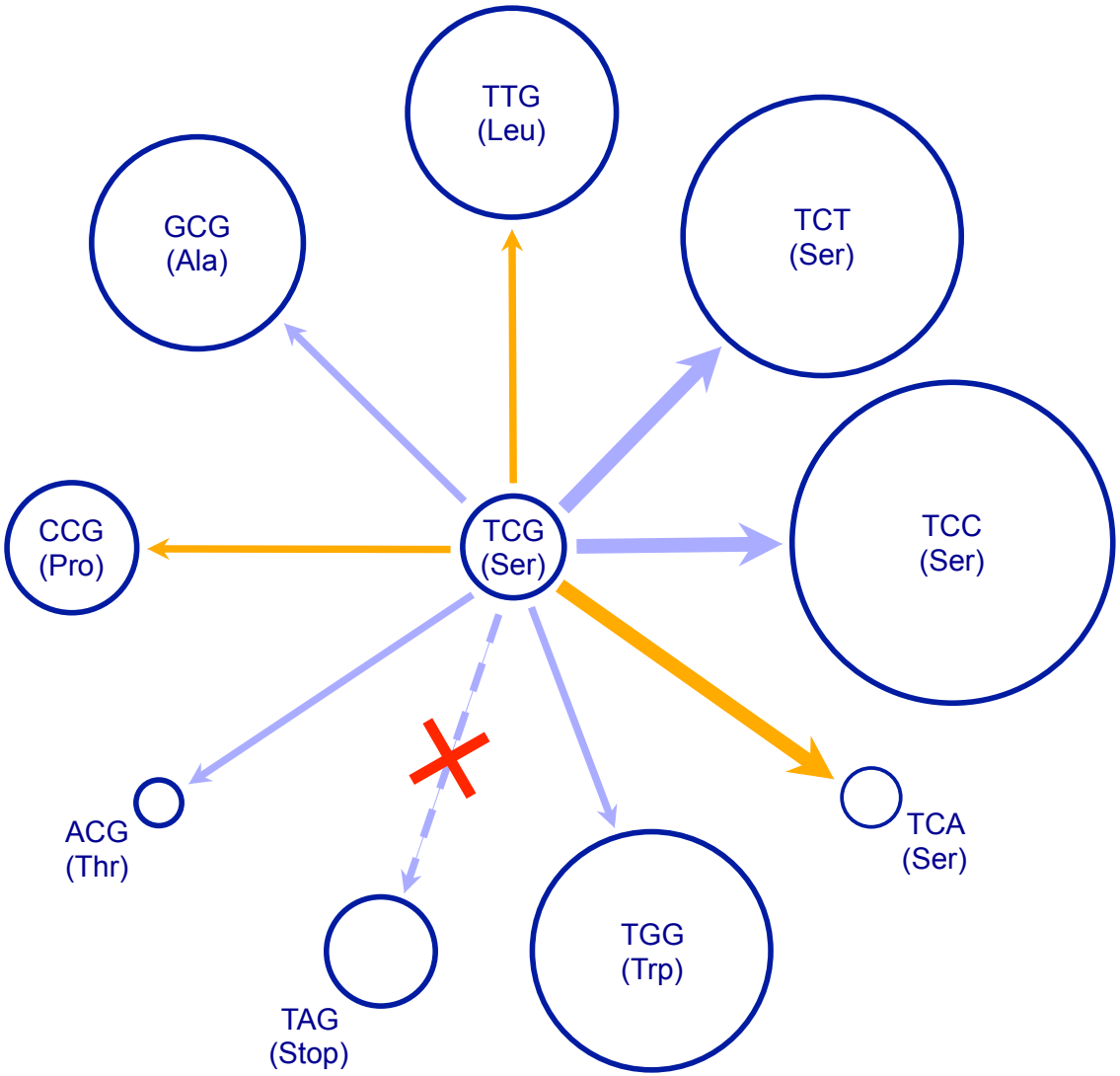
Approches au maximum de vraisemblance

- Publication la même année de deux méthodes similaires :
  - Goldman et Yang (1994) :
    - Simplification ultérieure par Yang et Nielsen (2000).
  - Muse et Gaut (1994).
- Dans les deux cas, définition d'un modèle de substitution spécifique aux codons.

Modèle de Goldman et Yang

- Indépendance des sites à l'intérieur d'un codon.
- À chaque intervalle de temps dt, une seule des trois positions est susceptible de muter.
- À chaque instant t, tout codon est susceptible de se substituer vers l'un de ses voisins :
  - Un voisin est un codon différant du codon d'intérêt par une seule substitution :
    - Position I, II ou III.
    - Codons Stop non considérés.
  - Chaque codon possède au plus neuf voisins.

Exemple du codon TCG^{Ser}



Paramètres du modèle

- Matrice $\mathbf{Q} = (q_{ij})$ des taux instantanés ($i, j \in \{\text{AAA, AAC, AAG, } \ldots, \text{TTT}\}$, codons Stop exclus) :

$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\text{AAA}} & q_{\text{AAA,AAC}} & \cdots & q_{\text{AAA,TTT}} \\ q_{\text{AAC,AAA}} & -\lambda_{\text{AAC}} & \cdots & q_{\text{AAC,TTT}} \\ \vdots & \vdots & \ddots & \vdots \\ q_{\text{TTT,AAA}} & q_{\text{TTT,AAC}} & \cdots & -\lambda_{\text{TTT}} \end{pmatrix}$$

- Ratio des taux de transitions/transversions κ.
- Ratio des distances non synonymes/synonymes $\omega = d_{\text{N}}/d_{\text{S}}$.

Taux de substitutions instantanés

- Les valeurs de q_{ij} sont telles que :

$$q_{ij} = \begin{cases} 0, & \text{si } i \text{ et } j \text{ diffèrent en plus d'une position} \\ \pi_j, & \text{pour une transversion synonyme} \\ \kappa\pi_j, & \text{pour une transition synonyme} \\ \omega\pi_j, & \text{pour une transversion non synonyme} \\ \omega\kappa\pi_j, & \text{pour une transition non synonyme} \end{cases}$$

- Normalisation de telle façon que le taux moyen de substitutions soit égal à un :

$$\sum_i \pi_i \lambda_i = 1$$

sachant que $\lambda_i = \sum_{j \neq i} q_{ij}$.

Probabilités de transition

- Comme pour les modèles standards, les valeurs des probabilités de transition $p_{ij}(t)$ sont données en résolvant $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$.
- À partir des valeurs de $p_{ij}(t)$, calcul des valeurs de $f_{ij}(t)$:

$$f_{ij}(t) = \pi_i p_{ij}(t)$$

  soit la probabilité d'observer le codon i de la séquence A aligné avec le codon j de la séquence B.
- Valeurs de π_i :
  - Uniforme ($\pi_i = 1/61, \forall i$).
  - À partir des fréquences des 61 codons dans le jeu de données (F61).
  - À partir des fréquences des nucléotides, toutes positions confondues (F1×4).
  - À partir des fréquences des nucléotides à chacune des trois positions des codons (F3×4).

Modélisation de l'hétérogénéité

- Dans une phylogénie, certaines lignées peuvent être soumises à de la sélection et d'autres non.
- Utilisation de plusieurs modèles afin de pouvoir détecter ces phénomènes :
  - Même valeur de ω pour toutes les branches de l'arbre (homogénéité).
  - Autant de valeurs de ω qu'il existe de branches dans l'arbre (hétérogénéité maximale).
  - Plusieurs intermédiaires entre ces deux extrêmes.
- Comparaison des différents modèles au moyen du LRT (*cf.* cours M. Gouy) afin de déterminer quel est le scénario le plus vraisemblable.

Séquences protéiques

- Premières séquences biologiques à avoir été utilisées pour construire des phylogénies moléculaires.
- Toujours fréquemment utilisées :
  - Plus conservées que les séquences d'ADN (substitutions synonymes) :
    - Utiles pour des analyses portant sur de longues durées évolutives ou sur des séquences évoluant rapidement.
    - Généralement inutilisables dans le cas d'organismes trop proches.
- Existence de nombreux modèles permettant d'estimer le nombre de substitutions entre deux séquences.

Modèle de Poisson

- Introduit par Zuckerkandl et Pauling (1965).
- Correction la plus simple pour les séquences protéiques :
  - Modélisation par une distribution de Poisson.
- Hypothèses :
  - Tous les sites évoluent indépendemment et selon le même processus.
  - Toutes les substitutions sont équiprobales.
  - Le taux de réversion est négligeable.
- Calcul de la distance au moyen de la formule :

$$d = -\ln(1 - p)$$

Modèle GTR pour les protéines ?

- Matrice 20×20 des taux instantanés :

$$\mathbf{Q} = \mathbf{S\Pi} = \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{R}} s_{\mathrm{AR}} & \cdots & \pi_{\mathrm{V}} s_{\mathrm{AV}} \\ \pi_{\mathrm{A}} s_{\mathrm{AR}} & -\lambda_{\mathrm{R}} & \cdots & \pi_{\mathrm{V}} s_{\mathrm{RV}} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{\mathrm{A}} s_{\mathrm{AV}} & \pi_{\mathrm{R}} s_{\mathrm{RV}} & \cdots & -\lambda_{\mathrm{V}} \end{pmatrix}$$

  Soit 189 paramètres d'échangeabilité s_{ij} et 19 fréquences à l'équilibre π_i $(i, j \in \{\mathrm{A}, \mathrm{R}, \mathrm{N}, \ldots, \mathrm{V}\})$.
- Non directement utilisable entre deux séquences :
  - Pas assez de données pour permettre l'estimation d'un si grand nombre de paramètres.
  - Estimation à partir de jeux de données de grande taille, puis fixation des paramètres.

Modèles empiriques

- Pas de définition *formelle* des probabilités de transition.
- Estimation des $p_{ij}(t)$ à partir de la fréquence des substitutions estimées sur des ensembles de séquences alignées :

$$f_{ij}(t) = \pi_i p_{ij}(t)$$

- Valeurs de $f_{ij}(t)$:
  - Inférence par maximum de parcimonie :
    - PAM (*Point Accepted Mutation*, Dayhoff *et al.*, 1978).
    - JTT (Jones, Taylor et Thornton, 1992).
  - Inférence par maximum de vraisemblance :
    - WAG (Whelan et Goldman, 2001).
    - LG (Le et Gascuel, 2008).

Calcul de la distance évolutive I

- Construction de la matrice de référence $\mathbf{P}(0.01)$ par normalisation des $\hat{p}_{ij}(t)$:
  - Choix arbitraire permettant de construire une matrice pour laquelle il est supposé qu'aucune substitution multiple ne s'est produite :

$$d = \varphi = p = 0.01$$

  - Comment calculer d si $p \neq 0.01$?
- Matrice de transition $\mathbf{P}(d)$ obtenue par exponentiation de $\mathbf{P}(0.01)$:

$$\mathbf{P}(d) = \mathbf{P}(0.01)^{d \times 100}$$

Calcul de la distance évolutive II

- La distance d est égale à la puissance (divisée par 100) à laquelle il faut élever $\mathbf{P}(0.01)$ de façon à obtenir $\varphi = p$:
  - Calcul itératif par approximations successives jusqu'à convergence vers la bonne valeur.
  - Exemple pour $p = 0.17$:
    - $\mathbf{P}(0.01) = \mathbf{P}(0.01)^1 \Rightarrow \varphi = 0.01 < p$
    - $\mathbf{P}(0.17) = \mathbf{P}(0.01)^{17} \Rightarrow \varphi \simeq 0.15201 < p$
    - $\mathbf{P}(0.18) = \mathbf{P}(0.01)^{18} \Rightarrow \varphi \simeq 0.15993 < p$
    - $\mathbf{P}(0.19) = \mathbf{P}(0.01)^{19} \Rightarrow \varphi \simeq 0.16774 < p$
    - $\mathbf{P}(0.20) = \mathbf{P}(0.01)^{20} \Rightarrow \varphi \simeq 0.17556 > p$
    - $\mathbf{P}(0.195) = \mathbf{P}(0.01)^{19.5} \Rightarrow \varphi \simeq 0.17162 > p$
    - $\mathbf{P}(0.1925) = \mathbf{P}(0.01)^{19.25} \Rightarrow \varphi \simeq 0.16968 < p$
    - $\mathbf{P}(0.19375) = \mathbf{P}(0.01)^{19.375} \Rightarrow \varphi \simeq 0.17065 > p$
    - $\ldots$
    - $\mathbf{P}(0.19291) = \mathbf{P}(0.01)^{19.291} \Rightarrow \varphi \simeq 0.17 \simeq p \Rightarrow d \simeq 0.19291$

Approximation de Kimura

- Calcul rapide d'une distance PAM avec les ordinateurs d'aujourd'hui, mais pas au moment de la conception du modèle.
- Mise en place par Kimura (1983) d'une mesure permettant d'approximer cette distance :

$$d = -\ln(1 - p - 0.2p^2)$$

- Méthode simple et rapide, mais présentant deux inconvénients :
  - Pas de possibilité de prise en compte des fréquences à l'équilibre des séquences étudiées.
  - La précision de l'estimation diminue avec le degré de divergence entre les séquences ($p \leq 0.75$).

Matrices de substitution

- Utilisées par les programmes d'alignement et de recherche de similarités :
  - Différentes des matrices de transition utilisées pour la reconstruction phylogénétique.
- Calcul effectué à partir des matrices de transition $\mathbf{P}(d)$ pour des valeurs *fixées* de d (0.3, 1, 1.5, 2.5, etc.) :
  - Soit $\hat{p}_{ij}(d)$ la probabilité d'une transition $i \rightarrow j$ estimée avec $\mathbf{P}(d)$.
  - Chaque élément $\delta_{ij}(d)$ de la matrice de substitution correspondante est défini par :

$$\delta_{ij}(d) = 10 \log \left(\frac{\hat{p}_{ij}(d)}{\pi_j} \right)$$

avec arrondi à l'entier le plus proche.

Données pour les autres modèles

- Modèle JTT :
  - Utilisation de 16300 séquences totalisant 59190 substitutions.
  - Procédure de construction identique à PAM.
- Modèle WAG :
  - Utilisation de 3905 séquences provenant de 182 familles.
  - Utilisation du maximum de vraisemblance pour estimer les probabilités de transitions :
    - Prise en compte des substitutions multiples.
- Modèle LG :
  - Utilisation de 49637 séquences provenant de 3912 familles.
  - Prise en compte des différences de vitesse d'évolution entre les sites.

Écriture et utilisation

- Dans les publications récentes, indication des valeurs de $\mathbf{S}$ et $\mathbf{\Pi}$ plutôt que de celles de $\mathbf{P}(t)$ ou $\mathbf{Q}$:
  - Fait pour permettre le remplacement facile des valeurs de π_i fournies par le modèle, sachant que :

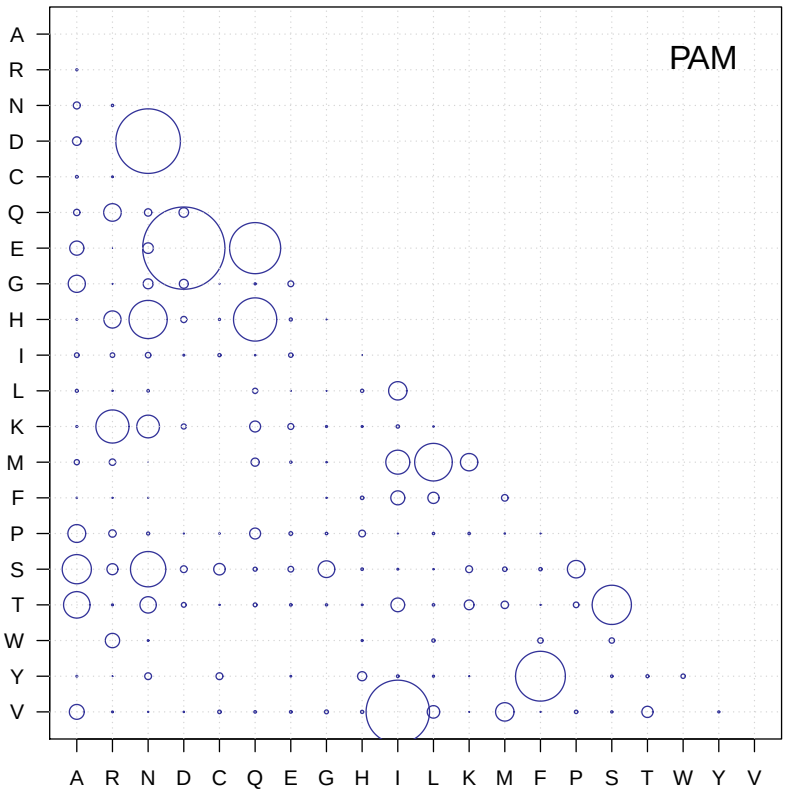
$$\mathbf{Q} = \mathbf{S\Pi}$$

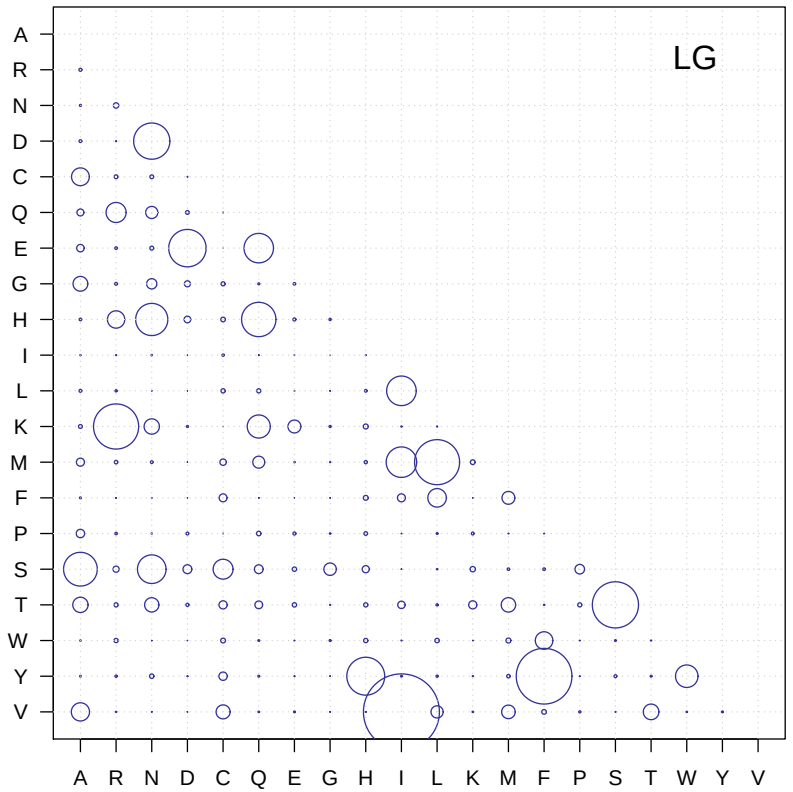
    avec la nécessité habituelle de normaliser les valeurs de q_{ij} de façon à ce que $\sum_i \pi_i \lambda_i = 1$.
  - Déduction des valeurs de $\mathbf{P}(t)$.
  - Calcul des distances évolutives avec la même procédure que celle utilisée pour PAM.

Comparaison des échangeabilités

Sur- ou sous-estimations de certaines valeurs de PAM :

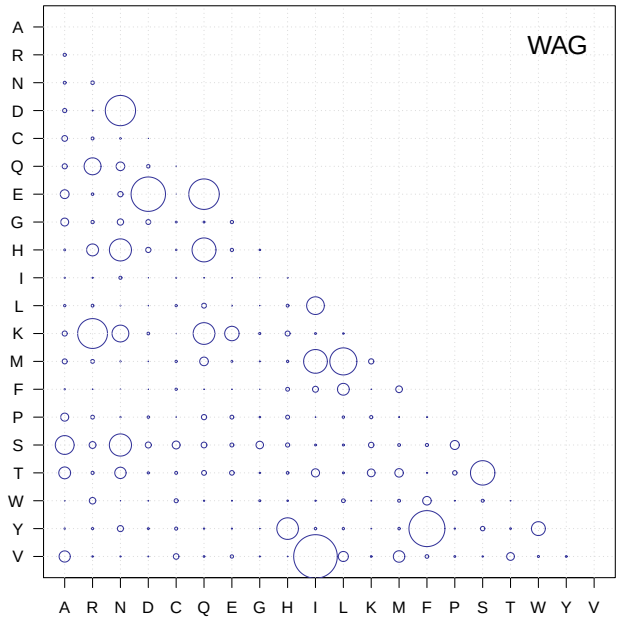
Problème lié à la taille de l'échantillon utilisé





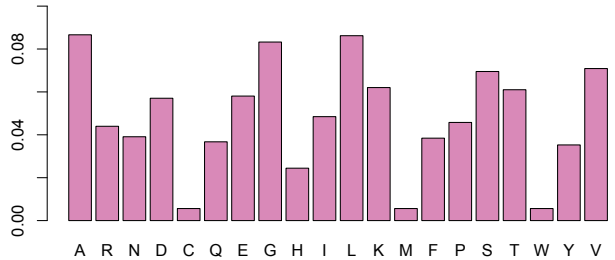
Approche classique

- Échangeabilités estimées à partir d'un jeu de données établi par les concepteurs du modèle.
- Fréquences à l'équilibre provenant du modèle ou obtenues à partir des séquences de l'alignement.



$\mathbf{S}$

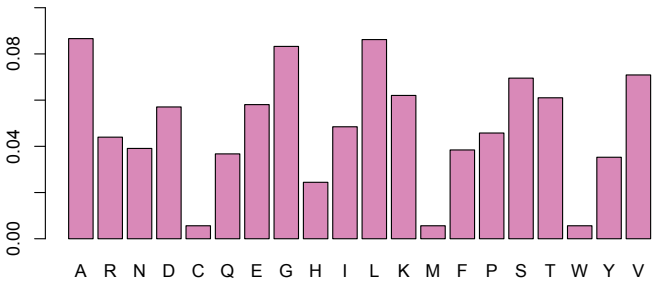
$\times$



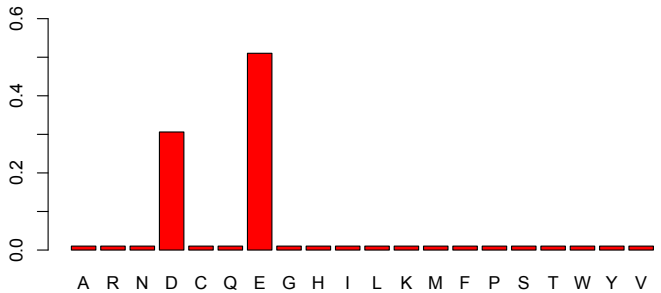
$\mathbf{\Pi}$

$= \mathbf{Q}$

Limites de l'approche classique



Fréquences à l'équilibre globales
(peu adaptées)

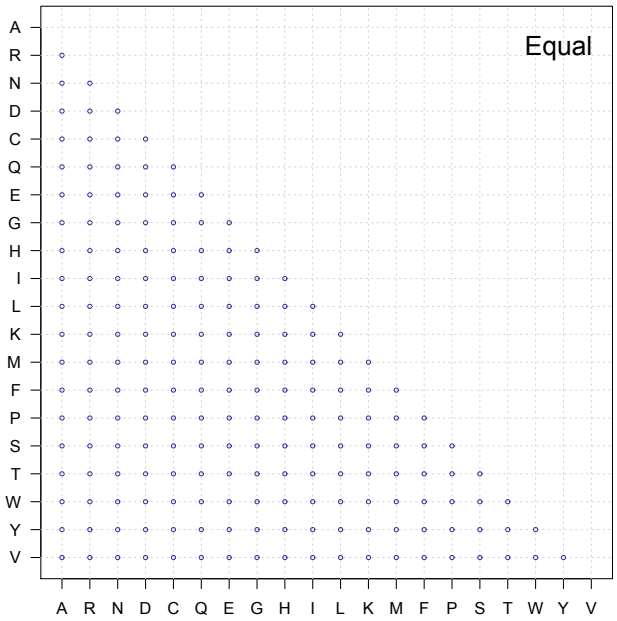


Fréquences à l'équilibre site spécifiques
(plus réalistes)

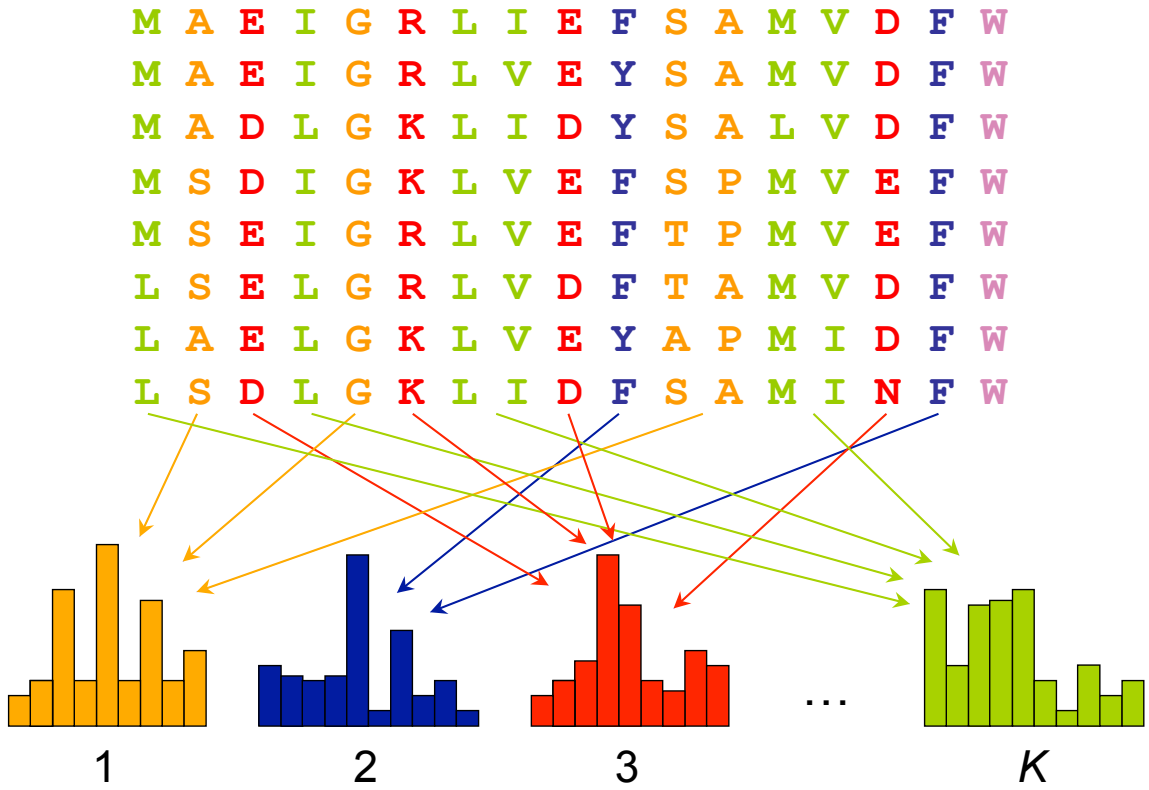
Approche site spécifique

- L'utilisation d'un jeu de valeurs π_i « globales » est non réaliste.
- Il n'est cependant pas possible d'utiliser un jeu par site de l'alignement :
  - Risques de surparamétrisation.
- Développement du modèle CAT (Le *et al.*, 2008) dans lequel il existe des *catégories* de sites :
  - Fréquences à l'équilibre :
    - Un jeu de valeurs de π_i par catégorie.
    - Cinq variantes à 20, 30, 40, 50 et 60 catégories.
  - Échangeabilités :
    - Une valeur unique, à l'image du modèle F81 (CAT-Poissson).
    - Valeurs provenant des modèles classiques (*e.g.*, CAT-JTT).
    - Valeurs estimées sur le jeu de données (CAT-GTR).

CAT-Poisson



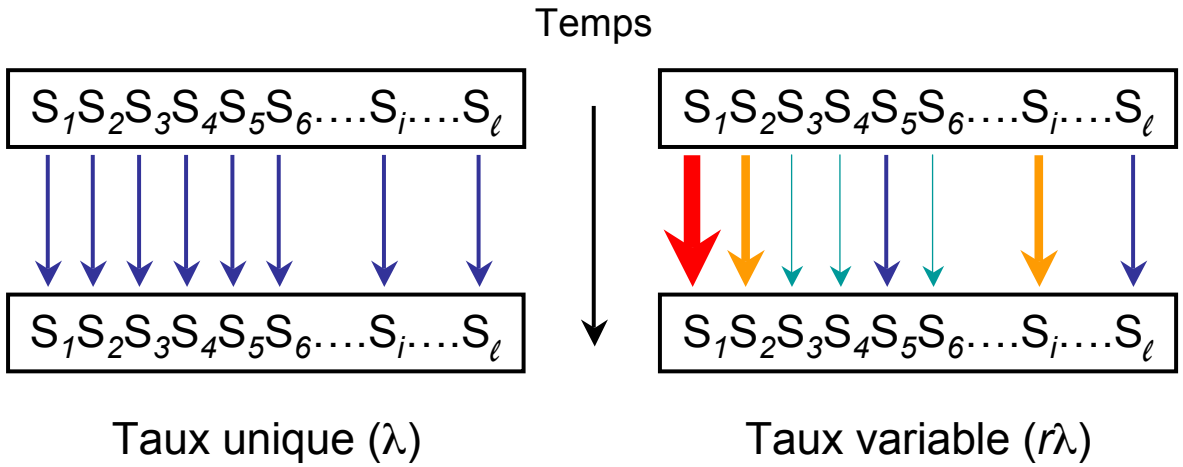
Une échangeabilité α



K catégories de valeurs de π_i
(K = 20, 30, 40, 50, 60)

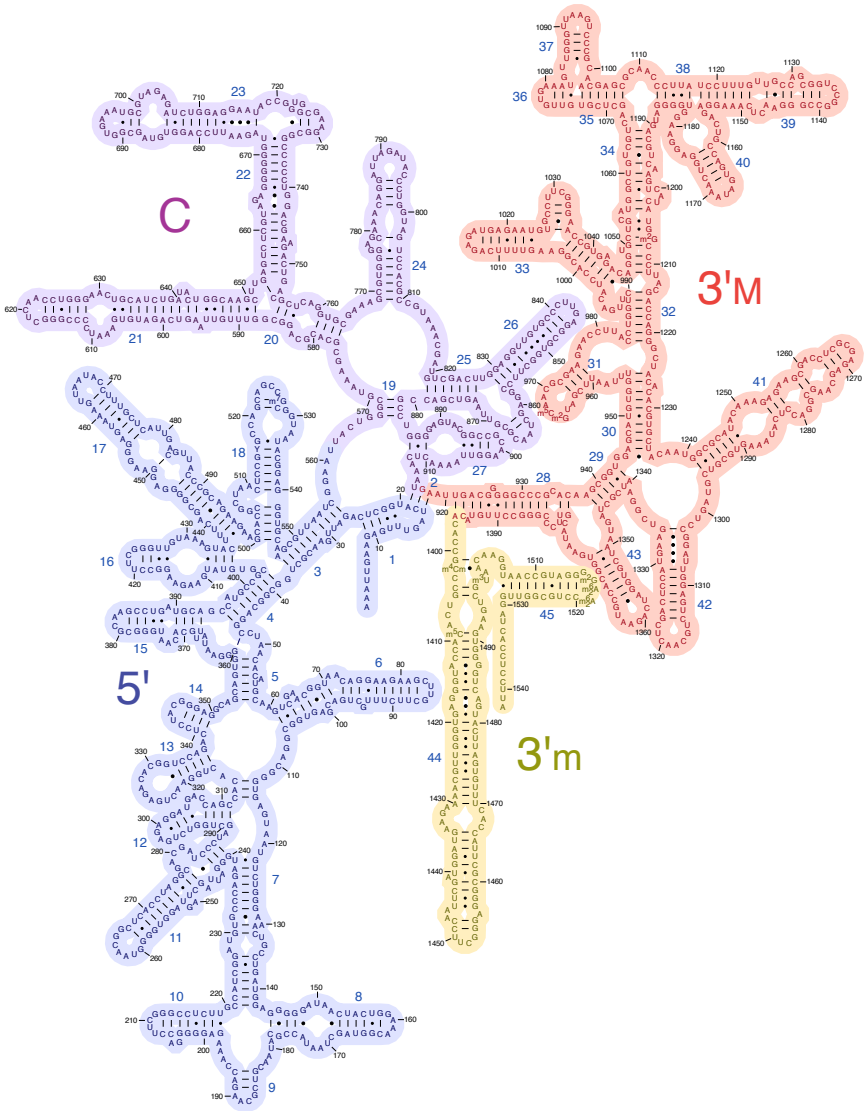
Vitesses variables

- Hypothèse des modèles « standards » :
  - Tous les sites possèdent le même taux instantané de substitution λ :
    - Existence de nombreux contre-exemples (séquences codantes, ARNr).
- Introduction d'un facteur correctif r :
  - Plusieurs méthodes de détermination possible :
    - Approche discrète simple, modèles mixtes finis, approche continue.



Exemple de l'ARNr 16S

- Marqueur couramment utilisé en phylogénie.
- Structure secondaire indispensable à la fonction.
- Taux d'évolutions différents suivant les régions :
  - Régions appariées évoluant lentement.
  - Régions dans les boucles évoluant rapidement.



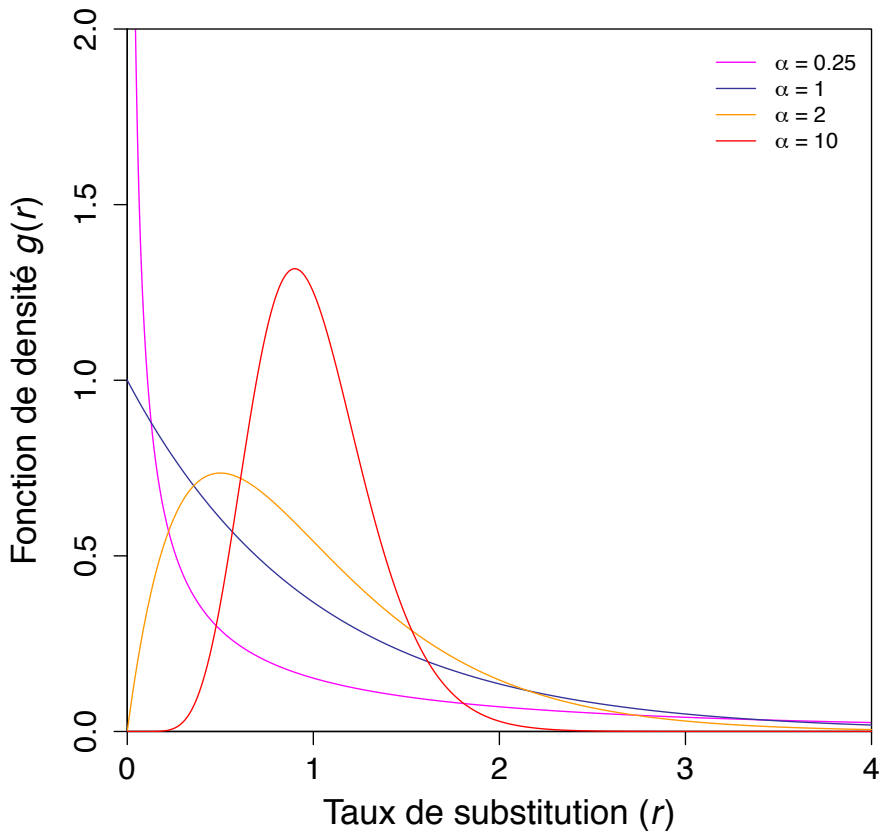
Correction par la loi Gamma

- Fonction de densité de probabilité $\mathcal{G}(\alpha, \beta)$ telle que :

$$g(r) = \frac{r^{\alpha-1} \mathrm{e}^{-r/\beta}}{\Gamma(\alpha)\beta^{\alpha}}$$

avec α le paramètre de *forme* et β le paramètre *d'échelle*.

- Détermination de α, avec $\beta = 1/\alpha$, de façon à ce que :
  - Moyenne : $\alpha\beta = 1$.
  - Variance : $\alpha\beta^2 = 1/\alpha$.



Discrétisation

- Nombre de classes fixé par l'utilisateur ($2 \leq K \leq 8$).
- Bornes z_k ($k = \{1, 2, \ldots, K-1\}$) calculées de façon à ce que le taux d'un site tiré au hasard ait une probabilité $1/K$ d'appartenir à chacune d'entre elles.
- Ajout éventuel d'une classe supplémentaire pour prendre en compte les sites *invariants*.
- Détermination du taux r_k associé à une classe par résolution du produit des intégrales :

$$r_k = \int_{z_i}^{z_j} rg(r)\mathrm{d}r \int_{z_i}^{z_j} g(r)\mathrm{d}r$$

pour la classe k délimitée par les bornes z_i et z_j.

Notations usuelles

- Indication des corrections éventuellement apportées à la version « standard » des modèles.
- Exemple avec le modèle PAM :
  - Si estimation des fréquences à l'équilibre en utilisant les séquences du jeu de données étudié : PAM-F.
  - Si, en plus du précédent, correction par une loi Gamma avec K classes : PAM-F+Γ_K.
  - Si, en plus du précédent, utilisation des invariants : PAM-F+Γ_K+I.
  - Toutes les combinaisons des trois modifications ci-dessus étant possibles.