Maximum de vraisemblance

Guy Perrière & Héloïse Philippon

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

31 novembre 2017

Historique

- Bases mathématiques développées dans les années 1920 par R.A. Fisher :
  - Génération d'estimateurs applicables à des cas plus complexes que ceux traités jusqu'alors en statistiques.
- Première application à la phylogénie moléculaire par Neyman (1971).
- Élargissement par Kashyap et Subas (1974) puis par Felsenstein (1981).
- Permet d'inférer des états de caractères ancestraux.
- Nécessite en théorie l'exploration de l'ensemble des topologies possibles.

Distribution discrète

- La *fonction de vraisemblance* d'une hypothèse H est définie par :

$$L(H) = \mathbb{P}(D|H)$$

soit la probabilité d'observer les données D sous l'hypothèse H.

- Maintenant, si D se décompose en ℓ observations indépendantes $D^{(i)}$ $(1 \leq i \leq \ell)$, alors :

$$\begin{aligned} L(H) &= \mathbb{P}(D^{(1)}|H) \times \mathbb{P}(D^{(2)}|H) \times \cdots \times \mathbb{P}(D^{(\ell)}|H) \\ &= \prod_{i=1}^{\ell} L^{(i)}(H) = \prod_{i=1}^{\ell} \mathbb{P}(D^{(i)}|H) \end{aligned}$$

Soit, sous forme logarithmique :

$$\ln L(H) = \sum_{i=1}^{\ell} \ln L^{(i)}(H) = \sum_{i=1}^{\ell} \ln \mathbb{P}(D^{(i)}|H)$$

Distribution continue

- Expression sous la forme d'une *fonction de densité*.
- Soit $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_\ell)$ un échantillon provenant d'une distribution de paramètres $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ inconnus.
- Dans ce cas, la fonction de vraisemblance associée est telle que :

$$L(\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \times f(x_2|\boldsymbol{\theta}) \times \cdots \times f(x_\ell|\boldsymbol{\theta})$$
$$= \prod_{i=1}^{\ell} f(x_i|\boldsymbol{\theta})$$

Soit, sous forme logarithmique :

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{\ell} \ln f(x_i|\boldsymbol{\theta})$$

Caractéristiques

- Maximiser la vraisemblance consiste à :
  - Trouver un ensemble d'estimations des paramètres $\hat{\boldsymbol{\theta}}$ de façon à ce que que $f(\mathbf{x}|\hat{\boldsymbol{\theta}})$ soit maximisée.
  - La fonction de vraisemblance $f(\mathbf{x}|\boldsymbol{\theta})$ n'est *pas* une fonction de densité de probabilité et, la plupart du temps :

$$\int f(\mathbf{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \neq 1$$

- Les estimations au maximum de vraisemblance sont :
  - Non biaisées ($\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$).
  - Consistantes (l'estimation converge vers la vraie valeur quand $\ell \to \infty$).
  - De variance minimale.

Notations pour la phylogénie

- En phylogénie moléculaire, les données sont représentées par un ensemble de séquences alignées S :
  - Chaque site dans l'alignement est désigné par le terme $S^{(i)}$ $(1 \leq i \leq \ell)$.
- Par ailleurs, le vecteur des paramètres est $\boldsymbol{\theta} = (\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha)$, avec :
  - τ la topologie de l'arbre.
  - $\mathbf{b}$ le vecteur des longueurs de branches.
  - $\boldsymbol{\vartheta}$ le vecteur des paramètre du modèle d'évolution utilisé.
  - α le paramètre de forme de la loi Gamma, le cas échéant.
- On en déduit l'expression de la vraisemblance de S, étant donné $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = \mathbb{P}(S|\boldsymbol{\theta}) = \prod_{i=1}^{\ell} \mathbb{P}\Big(S^{(i)}|\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha\Big)$$

Modèle de Jukes et Cantor

- Le calcul de la distance évolutive entre deux séquences au moyen du modèle de Jukes et Cantor est donnée par la formule :

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \Leftrightarrow p = \frac{3}{4} - \frac{3}{4}\mathrm{e}^{-4d/3}$$

- Soit ℓ le nombre de sites dans l'alignement et n le nombre de sites pour lesquels il y a une substitution entre les deux séquences :
  - Dans ce cas, la fonction de vraisemblance pour d est donnée par la loi binomiale $\mathcal{B}(\ell, p)$ telle que :

$$\begin{aligned} L(d) = f(p|d) &= \binom{\ell}{n} p^n (1-p)^{\ell-n} \\ &= \frac{\ell!}{n!(\ell-n)!}\left(\frac{3}{4} - \frac{3}{4}\mathrm{e}^{-4d/3}\right)^n \left(\frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4d/3}\right)^{\ell-n} \end{aligned}$$

Simplification des calculs

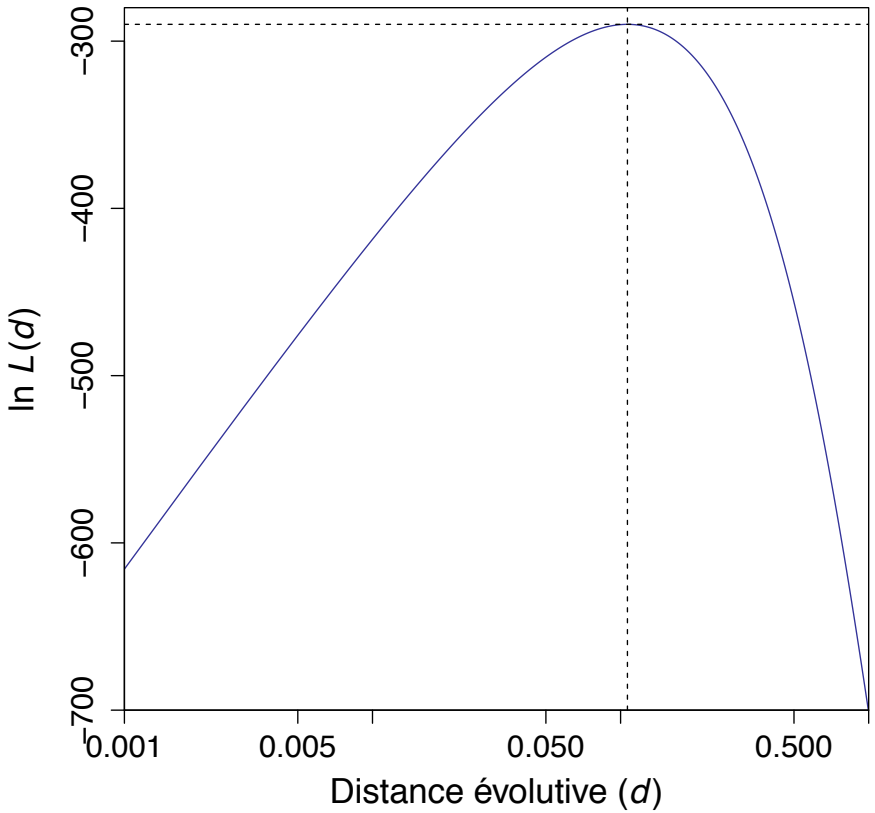
- Le coefficient binomial $\binom{\ell}{n}$ étant une constante, il peut être omis pour effectuer les calculs :
  - La vraisemblance obtenue change, mais le maximum sera toujours obtenu pour la même valeur de d.
- Passage en logarithmes pour éviter les dérives numériques du fait que les valeurs attendues sont très faibles :

$$\ln L(d) \propto n \ln \left(\frac{3}{4} - \frac{3}{4} \mathrm{e}^{-4d/3} \right) + (\ell - n) \ln \left(\frac{1}{4} + \frac{3}{4} \mathrm{e}^{-4d/3} \right)$$

- Variation des valeurs de d sur l'intervalle $[0.001, 2]$, réaliste du point de vue évolutif.

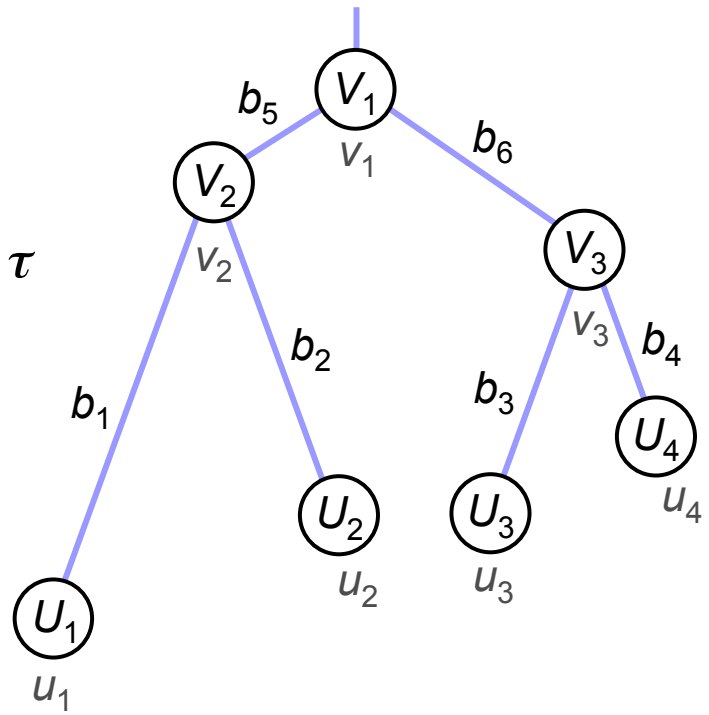
Application numérique

- Paire Homme-Gorille du jeu de données de Brown *et al.* (1982) :
  - $\ell = 896$
  - $n = 89$
- Calcul direct de la distance :
  - $d \simeq 0.1066$
- Estimation au maximum de vraisemblance :
  - $\max[\ln(L(d)] \simeq -289.95$ soit $d \simeq 0.1066$



Arbre à quatre UTO

- Soit un arbre à quatre UTO de topologie τ et dont les longueurs de branches sont fixées.
- U_1, U_2, U_3 et U_4 représentent les feuilles de l'arbre.
- V_1, V_2 et V_3 représentent les nœuds internes.
- Les états de caractères correspondants sont dénotés par u_1, u_2, u_3, u_4, v_1, v_2, $v_3 \in \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$.



Fonction de vraisemblance

- La vraisemblance à un site $S^{(i)}$ de l'alignement est telle que :

$$L^{(i)}(\boldsymbol{\theta}) = \mathbb{P}\Big(S^{(i)}|\tau, \mathbf{b}, \boldsymbol{\vartheta}\Big)$$
$$= \mathbb{P}(u_1, u_2, u_3, u_4, v_1, v_2, v_3|\tau, \mathbf{b}, \boldsymbol{\vartheta})$$

- Or les états ancestraux v_1, v_2 et v_3 sont inconnus :
  - Nécessité de prendre en compte tous les scénarios évolutifs possibles à chaque nœud interne de l'arbre.
  - L'expression de la vraisemblance s'écrit alors comme :

$$L^{(i)}(\boldsymbol{\theta}) = \sum_{v_1}\sum_{v_2}\sum_{v_3} \mathbb{P}(v_1)\mathbb{P}(v_2|v_1, b_5)\mathbb{P}(v_3|v_1, b_6)\mathbb{P}(u_1|v_2, b_1)$$
$$\times \mathbb{P}(u_2|v_2, b_2)\mathbb{P}(u_3|v_3, b_3)\mathbb{P}(u_4|v_3, b_4)$$

Calcul de la vraisemblance

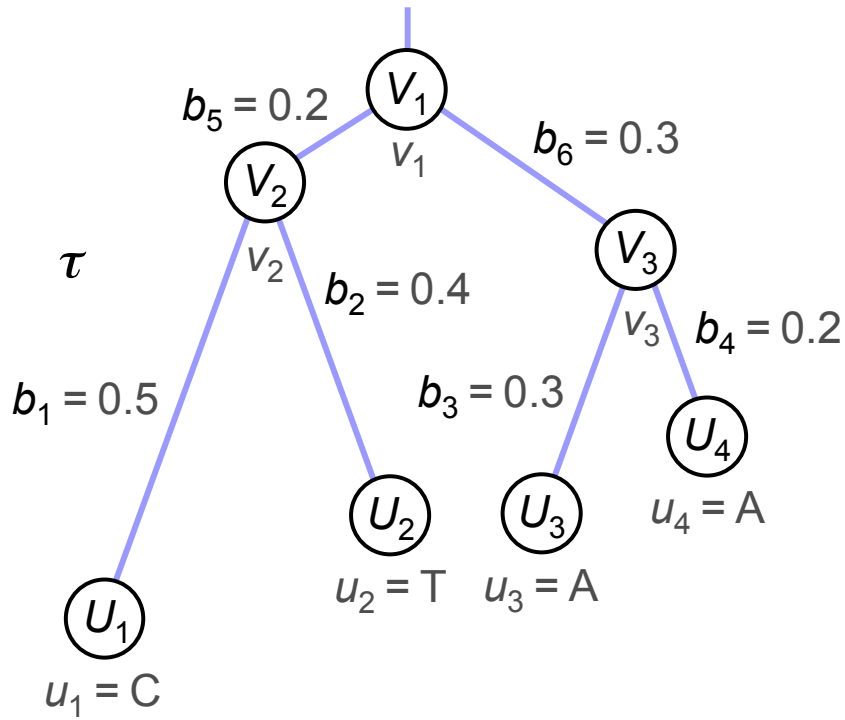
- La détermination de la vraisemblance totale nécessite le calcul de $L^{(i)}(\boldsymbol{\theta})$ pour chacun des ℓ sites.
- Le calcul des probabilités conditionnelles $\mathbb{P}(x|y, b)$ se fait par l'intermédiaire des modèles probabilistes vus précédemment.
- Sous l'hypothèse que le processus markovien modélisant l'évolution des séquences est à l'état stationnaire, on a :

$$\mathbb{P}(v_1) = \pi_{v_1}$$

La valeur de π_{v_1} étant estimée par la fréquence de l'état de caractère v_1 dans S.

Exemple de calcul

- Données :
  - Site de l'alignement tel que : $u_1 = \mathrm{C}$, $u_2 = \mathrm{T}$, $u_3 = \mathrm{A}$, $u_4 = \mathrm{A}$.
- Vecteur des paramètres $\boldsymbol{\theta}$:
  - Topologie τ racinée en V_1.
  - Vecteur $\mathbf{b}$ des longueurs de branches tel que : $b_1 = 0.5$, $b_2 = 0.4$, $b_3 = b_6 = 0.3$, $b_4 = b_5 = 0.2$
  - Modèle de Jukes et Cantor à un paramètre (p) :
    - Fréquences à l'équilibre $\pi_i = 1/4 \ \forall i$.



Probabilités de substitution

- Calcul des probabilités de substitution associées à chaque branche de longueur b au moyen de la relation :

$$p(b) = \frac{3}{4} - \frac{3}{4}\mathrm{e}^{-4b/3}$$

  On en déduit les valeurs de $p(b)$ pour les différentes longueurs de branches observées :
  - $p(b_1) = p(0.5) = 0.36$
  - $p(b_2) = p(0.4) = 0.31$
  - $p(b_3) = p(b_6) = p(0.3) = 0.25$
  - $p(b_4) = p(b_5) = p(0.2) = 0.18$
- Les probabilités de substitution p_{ij} $(i \neq j)$ sont toutes égales à $p(b)/3$.
- Les probabilités de conservations p_{ii} sont toutes égales à $1 - p(b)$.

Matrices de substitution

- On en déduit les matrices de substitution $\mathbf{P}(b)$ associées aux différentes longueurs de branches :
  - Valeurs utilisées pour calculer les probabilités conditionnelles $\mathbb{P}(x|y,b)$:

$$\mathbf{P}(0.5) = \begin{pmatrix} 0.64 & 0.12 & 0.12 & 0.12 \\ 0.12 & 0.64 & 0.12 & 0.12 \\ 0.12 & 0.12 & 0.64 & 0.12 \\ 0.12 & 0.12 & 0.12 & 0.64 \end{pmatrix} \quad \mathbf{P}(0.4) = \begin{pmatrix} 0.69 & 0.10 & 0.10 & 0.10 \\ 0.10 & 0.69 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.69 & 0.10 \\ 0.10 & 0.10 & 0.10 & 0.69 \end{pmatrix}$$

$$\mathbf{P}(0.3) = \begin{pmatrix} 0.75 & 0.08 & 0.08 & 0.08 \\ 0.08 & 0.75 & 0.08 & 0.08 \\ 0.08 & 0.08 & 0.75 & 0.08 \\ 0.08 & 0.08 & 0.08 & 0.75 \end{pmatrix} \quad \mathbf{P}(0.2) = \begin{pmatrix} 0.82 & 0.06 & 0.06 & 0.06 \\ 0.06 & 0.82 & 0.06 & 0.06 \\ 0.06 & 0.06 & 0.82 & 0.06 \\ 0.06 & 0.06 & 0.06 & 0.82 \end{pmatrix}$$

Calcul d'une valeur

- On se place dans le cas où $v_1 = v_2 = v_3 = \text{A}$:
  - Calcul de :

$$\begin{aligned}
&\mathbb{P}(v_1 = \text{A})\mathbb{P}(v_2 = \text{A}|v_1 = \text{A}, b_5 = 0.2)\mathbb{P}(v_3 = \text{A}|v_1 = \text{A}, b_6 = 0.3) \\
&\times \mathbb{P}(u_1 = \text{C}|v_2 = \text{A}, b_1 = 0.5)\mathbb{P}(u_2 = \text{T}|v_2 = \text{A}, b_2 = 0.4) \\
&\times \mathbb{P}(u_3 = \text{A}|v_3 = \text{A}, b_3 = 0.3)\mathbb{P}(u_4 = \text{A}|v_3 = \text{A}, b_4 = 0.2)
\end{aligned}$$

Soit, avec une écriture simplifiée :

$$\begin{aligned}
&\mathbb{P}(\text{A})\mathbb{P}(\text{A}|\text{A}, 0.2)\mathbb{P}(\text{A}|\text{A}, 0.3)\mathbb{P}(\text{C}|\text{A}, 0.5)\mathbb{P}(\text{T}|\text{A}, 0.4) \\
&\times \mathbb{P}(\text{A}|\text{A}, 0.3)\mathbb{P}(\text{A}|\text{A}, 0.2) \\
&= \pi_{\text{A}} p_{\text{AA}}(0.2) p_{\text{AA}}(0.3) p_{\text{CA}}(0.5) p_{\text{TA}}(0.4) p_{\text{AA}}(0.3) p_{\text{AA}}(0.2) \\
&= 0.25 \times 0.82 \times 0.75 \times 0.12 \times 0.10 \times 0.75 \times 0.82 \\
&= 0.001134675
\end{aligned}$$

Calcul des toutes les combinaisons (I)

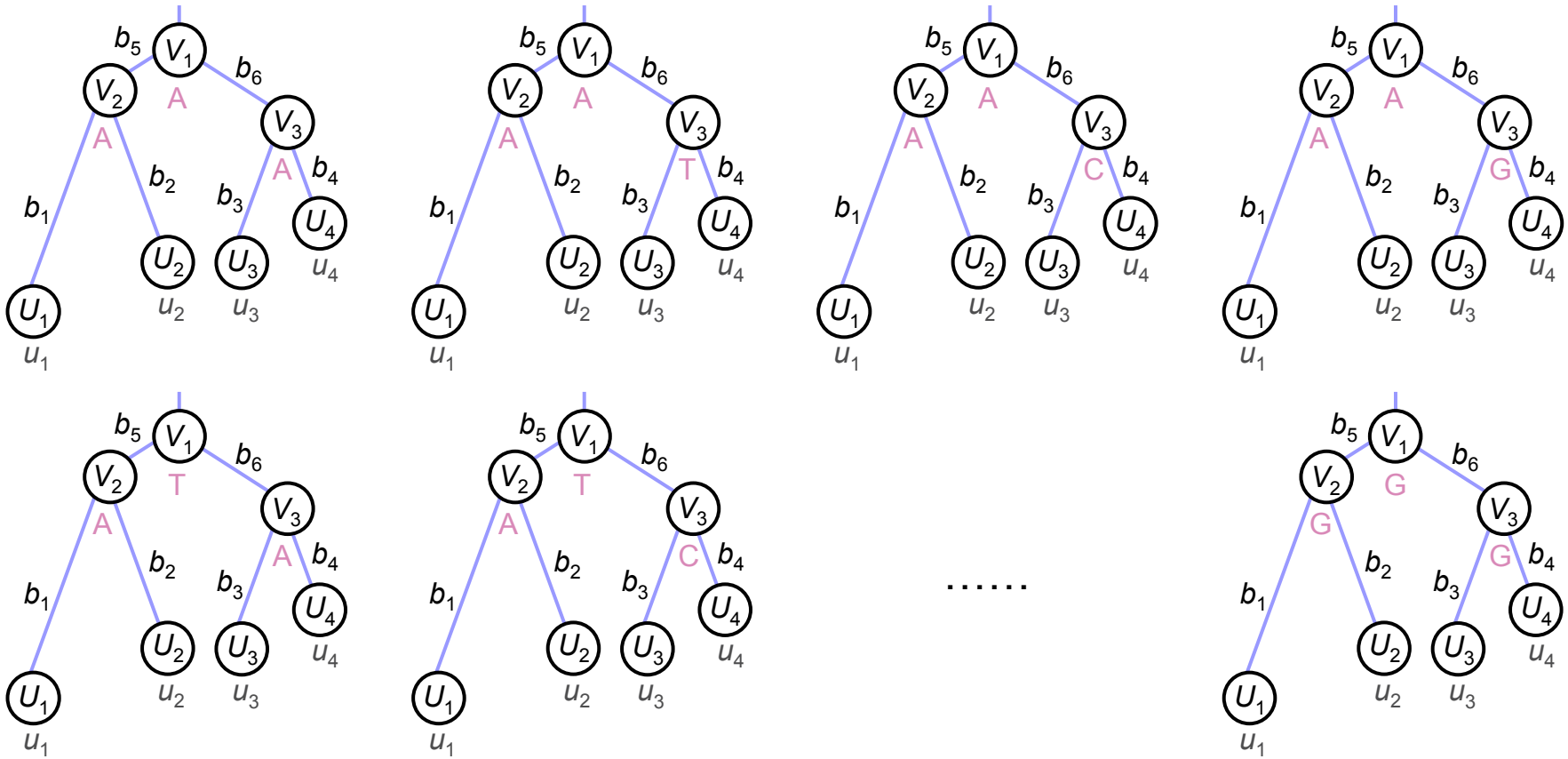
v_1	v_2	v_3	Vraisemblance	v_1	v_2	v_3	Vraisemblance
A	A	A	1.134675×10^{-3}	C	A	A	8.856×10^{-6}
A	A	C	9.4464×10^{-7}	C	A	C	6.48×10^{-7}
A	A	T	9.4464×10^{-7}	C	A	T	6.912×10^{-8}
A	A	G	9.4464×10^{-7}	C	A	G	6.912×10^{-8}
A	C	A	4.428×10^{-4}	C	C	A	6.45504×10^{-4}
A	C	C	3.6864×10^{-7}	C	C	C	4.7232×10^{-5}
A	C	T	3.6864×10^{-7}	C	C	T	5.03808×10^{-6}
A	C	G	3.6864×10^{-7}	C	C	G	5.03808×10^{-6}
A	T	A	5.728725×10^{-4}	C	T	A	6.11064×10^{-5}
A	T	C	4.76928×10^{-7}	C	T	C	4.4712×10^{-6}
A	T	T	4.76928×10^{-7}	C	T	T	4.76928×10^{-7}
A	T	G	4.76928×10^{-7}	C	T	G	4.76928×10^{-7}
A	G	A	8.3025×10^{-5}	C	G	A	8.856×10^{-6}
A	G	C	6.912×10^{-8}	C	G	C	6.48×10^{-7}
A	G	T	6.912×10^{-8}	C	G	T	6.912×10^{-8}
A	G	G	6.912×10^{-8}	C	G	G	6.912×10^{-8}

Calcul des toutes les combinaisons (II)

v_1	v_2	v_3	Vraisemblance	v_1	v_2	v_3	Vraisemblance
T	A	A	8.856×10^{-6}	G	A	A	8.856×10^{-6}
T	A	C	6.912×10^{-8}	G	A	C	6.912×10^{-8}
T	A	T	6.48×10^{-7}	G	A	T	6.912×10^{-8}
T	A	G	6.912×10^{-8}	G	A	G	6.48×10^{-7}
T	C	A	4.7232×10^{-5}	G	C	A	4.7232×10^{-5}
T	C	C	3.6864×10^{-7}	G	C	C	3.6864×10^{-7}
T	C	T	3.456×10^{-6}	G	C	T	3.6864×10^{-7}
T	C	G	3.6864×10^{-7}	G	C	G	3.456×10^{-6}
T	T	A	8.351208×10^{-4}	G	T	A	6.11064×10^{-5}
T	T	C	6.518016×10^{-6}	G	T	C	4.76928×10^{-7}
T	T	T	6.11064×10^{-5}	G	T	T	4.76928×10^{-7}
T	T	G	6.518016×10^{-6}	G	T	G	4.4712×10^{-6}
T	G	A	8.856×10^{-6}	G	G	A	1.21032×10^{-4}
T	G	C	6.912×10^{-8}	G	G	C	9.4464×10^{-7}
T	G	T	6.48×10^{-7}	G	G	T	9.4464×10^{-7}
T	G	G	6.912×10^{-8}	G	G	G	8.856×10^{-6}

Sommation de tous les termes : $L^{(i)}(\boldsymbol{\theta}) = 0.004267$

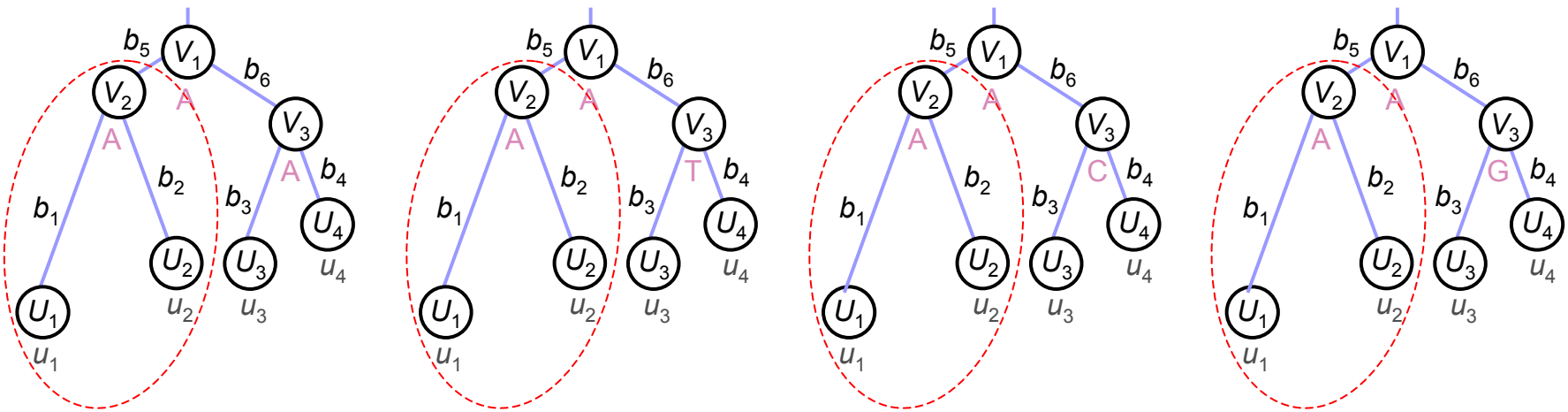
Ensemble des scénarios possibles



$4^3 = 64$ scénarios pour chaque site $S^{(i)}$

Nombre de termes de la fonction

- Le nombre de termes de la fonction de vraisemblance croît de façon exponentielle avec le nombre d'UTO :
  - Complexité en $O(\ell c^{n-1})$ pour le calcul de $L(\boldsymbol{\theta})$:
    - Avec $c = 4$ (séquences nucléotidiques) ou $c = 20$ (séquences protéiques).
  - Expression rapidement incalculable.
- Simplifications possibles, du fait que les mêmes valeurs sont recalculées de nombreuses fois :



Algorithme

- Felsenstein (1981) a proposé une méthode dite *d'élagage*, permettant de réduire très fortement la complexité des calculs :
  - Modification de la fonction de vraisemblance en décalant les sommations le plus à droite possible :

$$L^{(i)}(\boldsymbol{\theta}) = \sum_{v_1} \mathbb{P}(v_1) \left[\sum_{v_2} \mathbb{P}(v_2|v_1, b_5)\mathbb{P}(u_1|v_2, b_1)\mathbb{P}(u_2|v_2, b_2) \right] \times \left[\sum_{v_3} \mathbb{P}(v_3|v_1, b_6)\mathbb{P}(u_3|v_3, b_3)\mathbb{P}(u_4|v_3, b_4) \right]$$

  - Approche fondée sur le calcul de vraisemblances *conditionnelles* (ou *partielles*) $L_K^{(i)}(k)$ à chaque nœud K de l'arbre :
    - Probabilités d'observer les données aux feuilles du sous-arbre raciné par K, sachant l'état de caractère k à ce nœud.

Vraisemblances partielles d'une feuille

- Dans le cas de séquences nucléotidiques, si K correspond à une feuille quelconque de l'abre, alors :
  - $L_K^{(i)}(k) = 1$ pour l'un des quatre états de caractère et $L_K^{(i)}(k) = 0$ pour les trois autres ($k \in \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$).
  - Par exemple, si le nucléotide C est observé à la feuille U_1, alors le vecteur des vraisemblances partielles correspondant est :

$$\mathbf{L}_{U_1}^{(i)} = \left(L_{U_1}^{(i)}(\mathrm{A}), L_{U_1}^{(i)}(\mathrm{C}), L_{U_1}^{(i)}(\mathrm{T}), L_{U_1}^{(i)}(\mathrm{G})\right) = (0, 1, 0, 0)$$

  - Cette représentation permet de prendre en compte les ambigüités pouvant exister à certaines positions :
    - Pour une pyrimidine, le vecteur sera égal à $(0, 1, 1, 0)$.
    - Pour un *gap*, il sera égal à $(1, 1, 1, 1)$.

Vraisemblance partielle d'un nœud

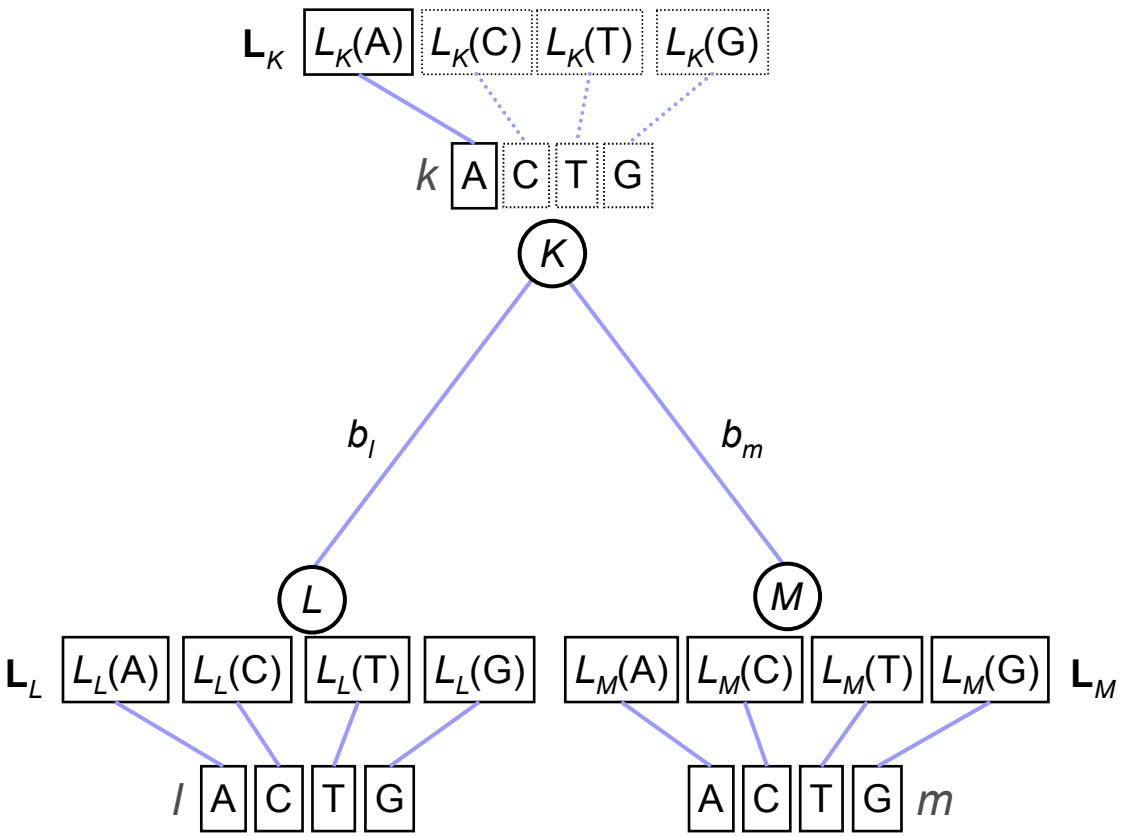
- Si K correspond à un nœud, alors :

$$L_K^{(i)}(k) = \sum_l \mathbb{P}(l|k, b_l) L_L^{(i)}(l) \times \sum_m \mathbb{P}(m|k, b_m) L_M^{(i)}(m)$$

  avec L et M les deux nœuds fils de K, b_l la longueur de la branche reliant K à L et b_m la longueur de la branche reliant K à M.
- En partant des feuilles, le calcul est réitéré jusqu'à atteindre la racine V_1 de l'arbre.
- À la racine, le vecteur des vraisemblances partielles obtenu permet de déterminer :

$$L^{(i)}(\boldsymbol{\theta}) = \sum_{v_1} \pi_{v_1} L_{V_1}^{(i)}(v_1)$$

Calcul à un nœud



Calcul de la vraisemblance partielle $L_K(A)$

Complexité de l'algorithme

- Aucune influence de la position de la racine sous l'hypothèse de réversibilité du processus markovien.
- Pour un site, c vraisemblances partielles sont déterminées pour chacun des $n-1$ nœuds de la topologie racinée :
  - Chacun de ces calculs implique le produit de deux termes, chaque terme étant le résultat d'une somme de c produits :
  - Complexité en $O(\ell n c^2)$ pour le calcul de $L(\boldsymbol{\theta})$:
    - Avec $c = 4$ (séquences nucléotidiques) ou $c = 20$ (séquences protéiques).
- Gains de temps possibles au moyen de certaines astuces :
  - Identification des sites identiques dans l'alignement afin d'éviter le recalcul de la même valeur.

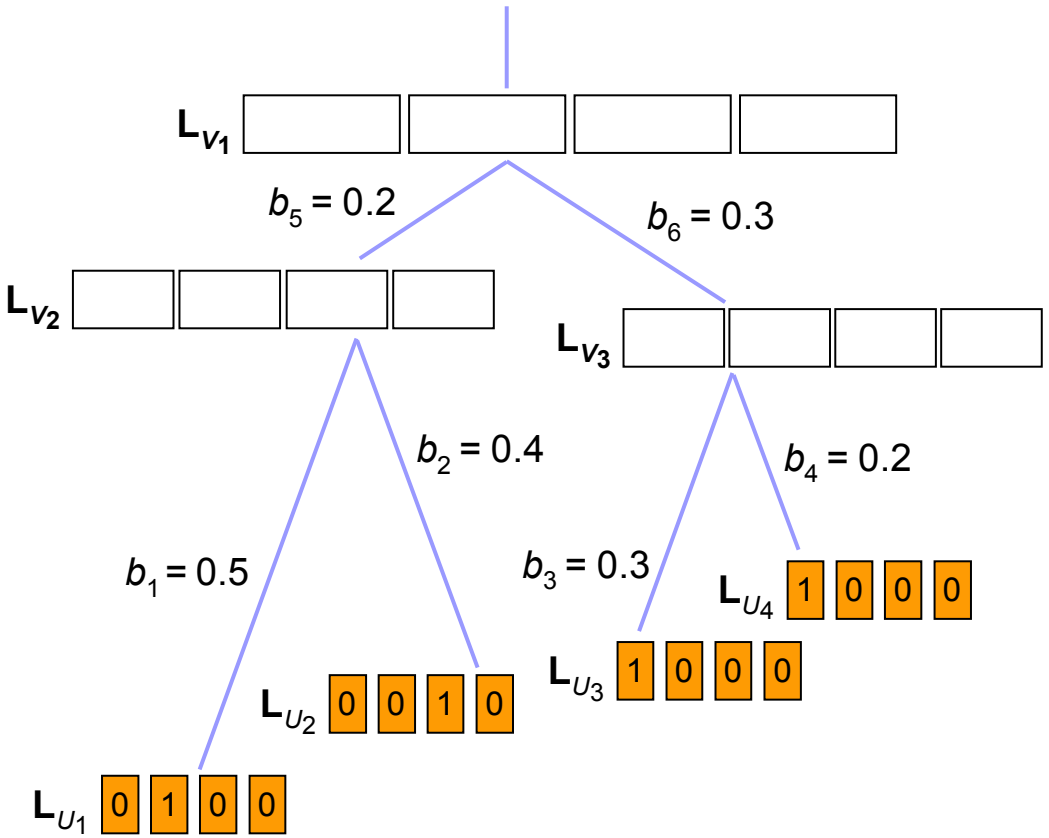
Vraisemblances partielles aux feuilles

- Sachant que $u_1 = \mathrm{C}$, $u_2 = \mathrm{T}$, $u_3 = \mathrm{A}$, $u_4 = \mathrm{A}$, les vecteurs de vraisemblances partielles aux feuilles sont donc tels que :

$$\mathbf{L}_{U_1}^{(i)} = \left(L_{U_1}^{(i)}(\mathrm{A}), L_{U_1}^{(i)}(\mathrm{C}), L_{U_1}^{(i)}(\mathrm{T}), L_{U_1}^{(i)}(\mathrm{G})\right) = (0, 1, 0, 0)$$
$$\mathbf{L}_{U_2}^{(i)} = \left(L_{U_2}^{(i)}(\mathrm{A}), L_{U_2}^{(i)}(\mathrm{C}), L_{U_2}^{(i)}(\mathrm{T}), L_{U_2}^{(i)}(\mathrm{G})\right) = (0, 0, 1, 0)$$
$$\mathbf{L}_{U_3}^{(i)} = \left(L_{U_3}^{(i)}(\mathrm{A}), L_{U_3}^{(i)}(\mathrm{C}), L_{U_3}^{(i)}(\mathrm{T}), L_{U_3}^{(i)}(\mathrm{G})\right) = (1, 0, 0, 0)$$
$$\mathbf{L}_{U_4}^{(i)} = \left(L_{U_4}^{(i)}(\mathrm{A}), L_{U_4}^{(i)}(\mathrm{C}), L_{U_4}^{(i)}(\mathrm{T}), L_{U_4}^{(i)}(\mathrm{G})\right) = (1, 0, 0, 0)$$

Vraisemblances partielles aux feuilles

- Initialisation du calcul de $L^{(i)}(\boldsymbol{\theta})$ aux feuilles :



Vraisemblances partielles au nœud V_2

- Calcul de $L^{(i)}_{V_2}(\mathrm{A})$:

$$
\begin{aligned}
L^{(i)}_{V_2}(\mathrm{A}) &= \left[p_{\mathrm{AA}}(0.5)L^{(i)}_{U_1}(\mathrm{A}) + p_{\mathrm{AC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + p_{\mathrm{AT}}(0.5)L^{(i)}_{U_1}(\mathrm{T}) + p_{\mathrm{AG}}(0.5)L^{(i)}_{U_1}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{AA}}(0.4)L^{(i)}_{U_2}(\mathrm{A}) + p_{\mathrm{AC}}(0.4)L^{(i)}_{U_2}(\mathrm{C}) + p_{\mathrm{AT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + p_{\mathrm{AG}}(0.4)L^{(i)}_{U_2}(\mathrm{G})\right] \\
&= \left[0 + p_{\mathrm{AC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + 0 + 0\right] \times \left[0 + 0 + p_{\mathrm{AT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + 0\right] \\
&= 0.12 \times 1 \times 0.10 \times 1 = 0.012
\end{aligned}
$$

- Calcul de $L^{(i)}_{V_2}(\mathrm{C})$:

$$
\begin{aligned}
L^{(i)}_{V_2}(\mathrm{C}) &= \left[p_{\mathrm{CA}}(0.5)L^{(i)}_{U_1}(\mathrm{A}) + p_{\mathrm{CC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + p_{\mathrm{CT}}(0.5)L^{(i)}_{U_1}(\mathrm{T}) + p_{\mathrm{CG}}(0.5)L^{(i)}_{U_1}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{CA}}(0.4)L^{(i)}_{U_2}(\mathrm{A}) + p_{\mathrm{CC}}(0.4)L^{(i)}_{U_2}(\mathrm{C}) + p_{\mathrm{CT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + p_{\mathrm{CG}}(0.4)L^{(i)}_{U_2}(\mathrm{G})\right] \\
&= \left[0 + p_{\mathrm{CC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + 0 + 0\right] \times \left[0 + 0 + p_{\mathrm{CT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + 0\right] \\
&= 0.64 \times 1 \times 0.10 \times 1 = 0.064
\end{aligned}
$$

Vraisemblances partielles au nœud V_2

- Calcul de $L^{(i)}_{V_2}(\mathrm{T})$:

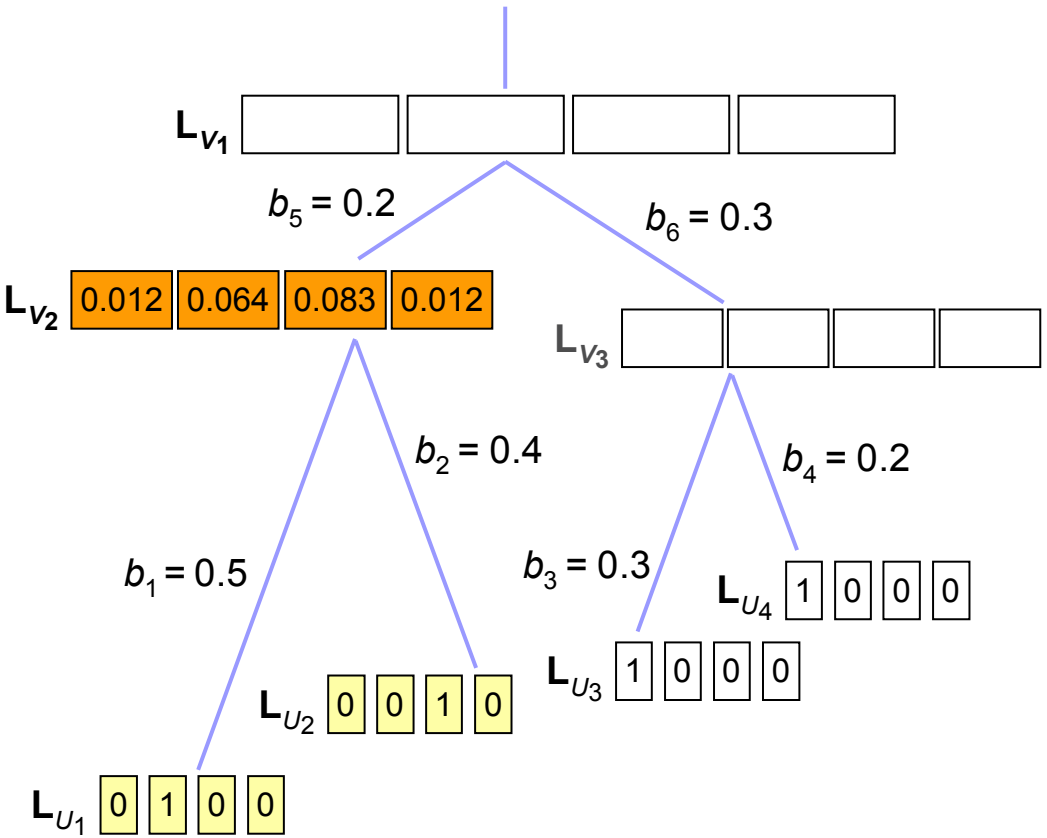
$$\begin{aligned}
L^{(i)}_{V_2}(\mathrm{T}) &= \left[p_{\mathrm{TA}}(0.5)L^{(i)}_{U_1}(\mathrm{A}) + p_{\mathrm{TC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + p_{\mathrm{TT}}(0.5)L^{(i)}_{U_1}(\mathrm{T}) + p_{\mathrm{TG}}(0.5)L^{(i)}_{U_1}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{TA}}(0.4)L^{(i)}_{U_2}(\mathrm{A}) + p_{\mathrm{TC}}(0.4)L^{(i)}_{U_2}(\mathrm{C}) + p_{\mathrm{TT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + p_{\mathrm{TG}}(0.4)L^{(i)}_{U_2}(\mathrm{G})\right] \\
&= \left[0 + p_{\mathrm{TC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + 0 + 0\right] \times \left[0 + 0 + p_{\mathrm{TT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + 0\right] \\
&= 0.12 \times 1 \times 0.69 \times 1 = 0.0828
\end{aligned}$$

- Calcul de $L^{(i)}_{V_2}(\mathrm{G})$:

$$\begin{aligned}
L^{(i)}_{V_2}(\mathrm{C}) &= \left[p_{\mathrm{GA}}(0.5)L^{(i)}_{U_1}(\mathrm{A}) + p_{\mathrm{GC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + p_{\mathrm{GT}}(0.5)L^{(i)}_{U_1}(\mathrm{T}) + p_{\mathrm{GG}}(0.5)L^{(i)}_{U_1}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{GA}}(0.4)L^{(i)}_{U_2}(\mathrm{A}) + p_{\mathrm{GC}}(0.4)L^{(i)}_{U_2}(\mathrm{C}) + p_{\mathrm{GT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + p_{\mathrm{GG}}(0.4)L^{(i)}_{U_2}(\mathrm{G})\right] \\
&= \left[0 + p_{\mathrm{GC}}(0.5)L^{(i)}_{U_1}(\mathrm{C}) + 0 + 0\right] \times \left[0 + 0 + p_{\mathrm{GT}}(0.4)L^{(i)}_{U_2}(\mathrm{T}) + 0\right] \\
&= 0.12 \times 1 \times 0.10 \times 1 = 0.012
\end{aligned}$$

Vraisemblances partielles au nœud V_2

- Construction du vecteur des vraisemblances partielles $\mathbf{L}_{V_2}^{(i)}$:



Vraisemblances partielles au nœud V_3

- Calcul de $L^{(i)}_{V_3}(\mathrm{A})$:

$$\begin{aligned}
L^{(i)}_{V_3}(\mathrm{A}) &= \left[p_{\mathrm{AA}}(0.3)L^{(i)}_{U_3}(\mathrm{A}) + p_{\mathrm{AC}}(0.3)L^{(i)}_{U_3}(\mathrm{C}) + p_{\mathrm{AT}}(0.3)L^{(i)}_{U_3}(\mathrm{T}) + p_{\mathrm{AG}}(0.3)L^{(i)}_{U_3}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{AA}}(0.2)L^{(i)}_{U_4}(\mathrm{A}) + p_{\mathrm{AC}}(0.2)L^{(i)}_{U_4}(\mathrm{C}) + p_{\mathrm{AT}}(0.2)L^{(i)}_{U_4}(\mathrm{T}) + p_{\mathrm{AG}}(0.2)L^{(i)}_{U_4}(\mathrm{G})\right] \\
&= \left[p_{\mathrm{AA}}(0.3)L^{(i)}_{U_3}(\mathrm{A}) + 0 + 0 + 0\right] \times \left[p_{\mathrm{AA}}(0.2)L^{(i)}_{U_4}(\mathrm{A}) + 0 + 0 + 0\right] \\
&= 0.75 \times 1 \times 0.82 \times 1 = 0.615
\end{aligned}$$

- Calcul de $L^{(i)}_{V_3}(\mathrm{C})$:

$$\begin{aligned}
L^{(i)}_{V_3}(\mathrm{C}) &= \left[p_{\mathrm{CA}}(0.3)L^{(i)}_{U_3}(\mathrm{A}) + p_{\mathrm{CC}}(0.3)L^{(i)}_{U_3}(\mathrm{C}) + p_{\mathrm{CT}}(0.3)L^{(i)}_{U_3}(\mathrm{T}) + p_{\mathrm{CG}}(0.3)L^{(i)}_{U_3}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{CA}}(0.2)L^{(i)}_{U_4}(\mathrm{A}) + p_{\mathrm{CC}}(0.2)L^{(i)}_{U_4}(\mathrm{C}) + p_{\mathrm{CT}}(0.2)L^{(i)}_{U_4}(\mathrm{T}) + p_{\mathrm{CG}}(0.2)L^{(i)}_{U_4}(\mathrm{G})\right] \\
&= \left[p_{\mathrm{CA}}(0.3)L^{(i)}_{U_3}(\mathrm{A}) + 0 + 0 + 0\right] \times \left[p_{\mathrm{CA}}(0.2)L^{(i)}_{U_4}(\mathrm{A}) + 0 + 0 + 0\right] \\
&= 0.08 \times 1 \times 0.06 \times 1 = 0.048
\end{aligned}$$

Vraisemblances partielles au nœud V_3

- Calcul de $L^{(i)}_{V_3}(\mathrm{T})$:

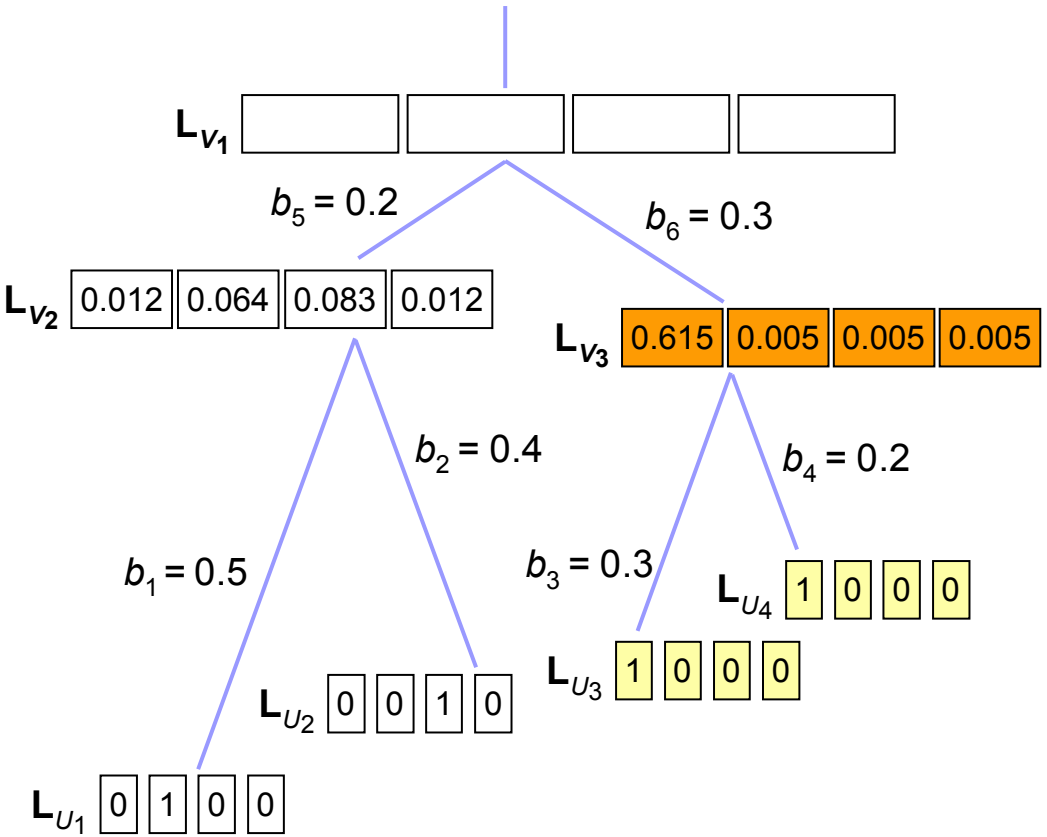
$$\begin{aligned}
L^{(i)}_{V_3}(\mathrm{T}) &= \left[p_{\mathrm{TA}}(0.3)L^{(i)}_{U_3}(\mathrm{A}) + p_{\mathrm{TC}}(0.3)L^{(i)}_{U_3}(\mathrm{C}) + p_{\mathrm{TT}}(0.3)L^{(i)}_{U_3}(\mathrm{T}) + p_{\mathrm{TG}}(0.3)L^{(i)}_{U_3}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{TA}}(0.2)L^{(i)}_{U_4}(\mathrm{A}) + p_{\mathrm{TC}}(0.2)L^{(i)}_{U_4}(\mathrm{C}) + p_{\mathrm{TT}}(0.2)L^{(i)}_{U_4}(\mathrm{T}) + p_{\mathrm{TG}}(0.2)L^{(i)}_{U_4}(\mathrm{G})\right] \\
&= \left[p_{\mathrm{TA}}(0.3)L^{(i)}_{U_3} + 0 + 0 + 0\right] \times \left[p_{\mathrm{TA}}(0.2)L^{(i)}_{U_4} + 0 + 0 + 0\right] \\
&= 0.08 \times 1 \times 0.06 \times 1 = 0.048
\end{aligned}$$

- Calcul de $L^{(i)}_{V_3}(\mathrm{G})$:

$$\begin{aligned}
L^{(i)}_{V_3}(\mathrm{C}) &= \left[p_{\mathrm{GA}}(0.3)L^{(i)}_{U_3}(\mathrm{A}) + p_{\mathrm{GC}}(0.3)L^{(i)}_{U_3}(\mathrm{C}) + p_{\mathrm{GT}}(0.3)L^{(i)}_{U_3}(\mathrm{T}) + p_{\mathrm{GG}}(0.3)L^{(i)}_{U_3}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{GA}}(0.2)L^{(i)}_{U_4}(\mathrm{A}) + p_{\mathrm{GC}}(0.2)L^{(i)}_{U_4}(\mathrm{C}) + p_{\mathrm{GT}}(0.2)L^{(i)}_{U_4}(\mathrm{T}) + p_{\mathrm{GG}}(0.2)L^{(i)}_{U_4}(\mathrm{G})\right] \\
&= \left[[p_{\mathrm{GA}}(0.3)L^{(i)}_{U_3} + 0 + 0 + 0\right] \times \left[p_{\mathrm{GA}}(0.2)L^{(i)}_{U_4} + 0 + 0 + 0\right] \\
&= 0.08 \times 1 \times 0.06 \times 1 = 0.048
\end{aligned}$$

Vraisemblances partielles au nœud V_3

- Construction du vecteur des vraisemblances partielles $\mathbf{L}_{V_3}^{(i)}$:



Vraisemblances partielles à la racine V_1

- Calcul de $L_{V_1}^{(i)}(\mathrm{A})$:

$$
\begin{aligned}
L_{V_1}^{(i)}(\mathrm{A}) &= \left[p_{\mathrm{AA}}(0.2)L_{V_2}^{(i)}(\mathrm{A}) + p_{\mathrm{AC}}(0.2)L_{V_2}^{(i)}(\mathrm{C}) + p_{\mathrm{AT}}(0.2)L_{V_2}^{(i)}(\mathrm{T}) + p_{\mathrm{AG}}(0.2)L_{V_2}^{(i)}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{AA}}(0.3)L_{V_3}^{(i)}(\mathrm{A}) + p_{\mathrm{AC}}(0.3)L_{V_3}^{(i)}(\mathrm{C}) + p_{\mathrm{AT}}(0.3)L_{V_3}^{(i)}(\mathrm{T}) + p_{\mathrm{AG}}(0.3)L_{V_3}^{(i)}(\mathrm{G})\right] \\
&= [0.82 \times 0.012 + 0.06 \times 0.064 + 0.06 \times 0.0828 + 0.06 \times 0.012] \\
&\times [0.75 \times 0.615 + 0.08 \times 0.0048 + 0.08 \times 0.0048 + 0.08 \times 0.0048] \\
&= 0.008956
\end{aligned}
$$

- Calcul de $L_{V_1}^{(i)}(\mathrm{C})$:

$$
\begin{aligned}
L_{V_1}^{(i)}(\mathrm{C}) &= \left[p_{\mathrm{CA}}(0.2)L_{V_2}^{(i)}(\mathrm{A}) + p_{\mathrm{CC}}(0.2)L_{V_2}^{(i)}(\mathrm{C}) + p_{\mathrm{CT}}(0.2)L_{V_2}^{(i)}(\mathrm{T}) + p_{\mathrm{CG}}(0.2)L_{V_2}^{(i)}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{CA}}(0.3)L_{V_3}^{(i)}(\mathrm{A}) + p_{\mathrm{CC}}(0.3)L_{V_3}^{(i)}(\mathrm{C}) + p_{\mathrm{CT}}(0.3)L_{V_3}^{(i)}(\mathrm{T}) + p_{\mathrm{CG}}(0.3)L_{V_3}^{(i)}(\mathrm{G})\right] \\
&= [0.06 \times 0.012 + 0.82 \times 0.064 + 0.06 \times 0.0828 + 0.06 \times 0.012] \\
&\times [0.08 \times 0.615 + 0.75 \times 0.0048 + 0.08 \times 0.0048 + 0.08 \times 0.0048] \\
&= 0.003155
\end{aligned}
$$

Vraisemblances partielles à la racine V_1

- Calcul de $L_{V_1}^{(i)}(\mathrm{T})$:

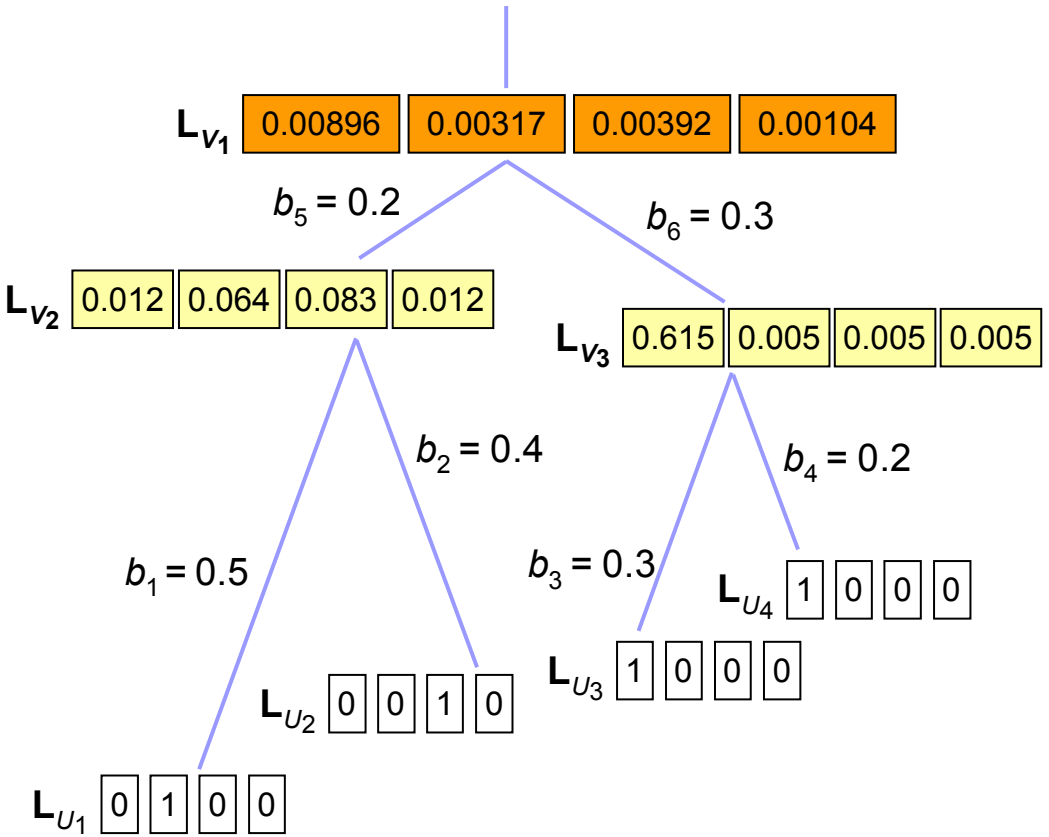
$$
\begin{aligned}
L_{V_1}^{(i)}(\mathrm{T}) &= \left[p_{\mathrm{TA}}(0.2)L_{V_2}^{(i)}(\mathrm{A}) + p_{\mathrm{TC}}(0.2)L_{V_2}^{(i)}(\mathrm{C}) + p_{\mathrm{TT}}(0.2)L_{V_2}^{(i)}(\mathrm{T}) + p_{\mathrm{TG}}(0.2)L_{V_2}^{(i)}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{TA}}(0.3)L_{V_3}^{(i)}(\mathrm{A}) + p_{\mathrm{TC}}(0.3)L_{V_3}^{(i)}(\mathrm{C}) + p_{\mathrm{TT}}(0.3)L_{V_3}^{(i)}(\mathrm{T}) + p_{\mathrm{TG}}(0.3)L_{V_3}^{(i)}(\mathrm{G})\right] \\
&= [0.06 \times 0.012 + 0.06 \times 0.064 + 0.82 \times 0.0828 + 0.06 \times 0.012] \\
&\times [0.08 \times 0.615 + 0.08 \times 0.0048 + 0.75 \times 0.0048 + 0.08 \times 0.0048] \\
&= 0.00392
\end{aligned}
$$

- Calcul de $L_{V_1}^{(i)}(\mathrm{G})$:

$$
\begin{aligned}
L_{V_1}^{(i)}(\mathrm{C}) &= \left[p_{\mathrm{GA}}(0.2)L_{V_2}^{(i)}(\mathrm{A}) + p_{\mathrm{GC}}(0.2)L_{V_2}^{(i)}(\mathrm{C}) + p_{\mathrm{GT}}(0.2)L_{V_2}^{(i)}(\mathrm{T}) + p_{\mathrm{GG}}(0.2)L_{V_2}^{(i)}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{GA}}(0.3)L_{V_3}^{(i)}(\mathrm{A}) + p_{\mathrm{GC}}(0.3)L_{V_3}^{(i)}(\mathrm{C}) + p_{\mathrm{GT}}(0.3)L_{V_3}^{(i)}(\mathrm{T}) + p_{\mathrm{GG}}(0.3)L_{V_3}^{(i)}(\mathrm{G})\right] \\
&= [0.06 \times 0.012 + 0.06 \times 0.064 + 0.06 \times 0.0828 + 0.82 \times 0.012] \\
&\times [0.08 \times 0.615 + 0.08 \times 0.0048 + 0.08 \times 0.0048 + 0.75 \times 0.0048] \\
&= 0.001038
\end{aligned}
$$

Vraisemblances partielles à la racine V_1

- Construction du vecteur des vraisemblances partielles $\mathbf{L}_{V_1}^{(i)}$:



Calcul de la vraisemblance au site $S^{(i)}$

- À partir du vecteur des vraisemblances partielles à la racine, on en déduit la valeur de $L^{(i)}(\boldsymbol{\theta})$:

$$\begin{aligned} L^{(i)}(\boldsymbol{\theta}) &= \sum_{v_1} \pi_{v_1} L^{(i)}_{V_1}(v_1) \\ &= \pi_{\mathrm{A}} L^{(i)}_{V_1}(\mathrm{A}) + \pi_{\mathrm{C}} L^{(i)}_{V_1}(\mathrm{C}) + \pi_{\mathrm{T}} L^{(i)}_{V_1}(\mathrm{T}) + \pi_{\mathrm{G}} L^{(i)}_{V_1}(\mathrm{G}) \\ &= \frac{1}{4}(0.008956 + 0.003155 + 0.00392 + 0.001038) \\ &= 0.004267 \end{aligned}$$

Soit, sous forme logarithmique :

$$\ln L^{(i)}(\boldsymbol{\theta}) = \ln(0.004267) \simeq -5.4568$$

Procédure générale

- En théorie, nécessité d'explorer l'ensemble des topologies et des combinaisons de longueurs de branches :
  - Impossible du fait de la croissance très rapide du nombre de topologies et du caractère continu des longueurs de branches.
- En pratique :
  - Exploration de l'espace des topologies via les heuristiques vues précédemment (NNI, SPR, TBR).
  - Optimisation branche par branche pour déterminer les longueurs maximisant la vraisemblance.
- Pour une topologie et un ensemble de longueurs de branches données :
  - Calcul des valeurs de vraisemblances par site $L^{(i)}(\boldsymbol{\theta})$:
    - Calcul de la vraisemblance globale $\ln L(\boldsymbol{\theta}) = \sum_i \ln L^{(i)}(\boldsymbol{\theta})$.

Avantages et limitations

- Méthode la mieux justifiée du point de vue théorique (si vous êtes fréquentiste).
- Donne de meilleurs résultats que la parcimonie ou les méthodes de distances dans la plupart des cas.
- Consistante si l'on utilise le bon modèle.
- Coûteuse en temps de calcul :
  - *Bootstrap* standard difficile d'utilisation.
- Risques de surparamétrisation avec les modèles trop complexes :
  - Tests pour sélectionner le modèle permettant d'obtenir le meilleur compromis vraisemblance/nombre de paramètres.

Performances en simulation

- Génération aléatoire de 5000 arbres à 40 UTO :
  - Variation des longueurs de branches.
- Construction des séquences d'ADN correspondantes :
  - Modèle de Kimura à deux paramètres.
- Qualité des reconstructions obtenues :
  - Distance topologique entre l'arbre vrai (connu) et l'arbre reconstruit.