Analyse de séquences et phylogénie moléculaire -

Maximum de vraisemblance

Ecole doctorale E2M2 - 2015-2016

(http://www.frangun.org)

Céline Brochier (celine.brochier-armanet@univ-lyon1.fr)
Guy Perrière (guy.perriere@univ-lyon1.fr)

Principe général

- Basé sur des lois de probabilité conditionnelles
- La vraisemblance de l'hypothèse H connaissant les données D est définie par:
  - $L = P(D|H)$ ⇔ probabilité d'observer les données D sous l'hypothèse H
  - ≠ $L = P(H|D)$: probabilité de l'hypothèse H sachant les données
  - Si on dispose de *n* observations indépendantes
    - $L = P(D^{(1)}|H) \times P(D^{(2)}|H) \times P(D^{(3)}|H) \times ... \times P(D^{(n)}|H)$

Un exemple simple

- Estimation de la *probabilité p* d'obtenir face d'une pièce lancée 11 fois
- Hypothèses :
  - Indépendance des lancés
  - Tous les lancés on la même probabilité *p* (inconnue) d'obtenir face
- Données : $L = P(D|p)$
  - Résultats observés : FFPPFPFFPPP
- Définition de la fonction de vraisemblance
- Soit

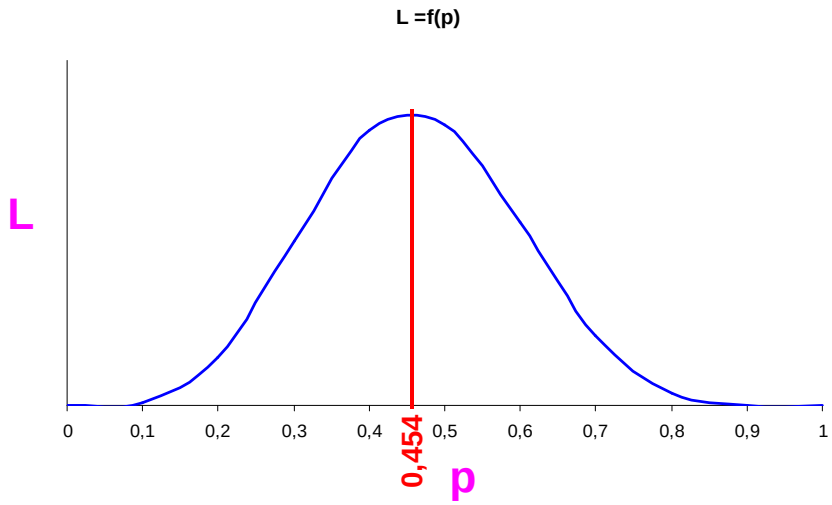
$$L = P(D|p) = p \times p \times (1-p) \times (1-p) \times p \times (1-p) \times p \times p \times (1-p) \times (1-p) \times (1-p)$$
$$= p^5 \times (1-p)^6$$

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Un exemple simple

- On cherche ensuite quelle est la valeur de p (parmi toutes les valeurs possibles) qui maximise la probabilité d'obtenir les données D, cad les résultats des lancers observés

$$L = P(D|p) = p \times p \times (1-p) \times (1-p) \times p \times (1-p) \times p \times p \times (1-p) \times (1-p) \times (1-p)$$
$$= p^5 \times (1-p)^6$$



La vraisemblance est maximale pour p = 0,454

Généralités - Maximum de Vraisemblance

- Introduit par Edwards et Cavalli-Sforza (1964) pour l'étude de données de type fréquences de gènes
- Appliquée à la phylogénie moléculaire par Neyman (1971)
- Élargissement par Kashyap et Subas (1974) et Felsenstein (1981)
- IDEE DE BASE
  - Étant donné un modèle d'évolution, on peut estimer une phylogénie avec des méthodes statistiques comme le maximum de vraisemblance

- PROPRIETES des estimations par Maximum de vraisemblance
  - Bonne consistance ⇔ convergent vers la valeur correcte du paramètre
  - Bonne efficience ⇔ ont la plus petite variance possible autour de la vraie valeur du paramètre

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Généralités - Maximum de Vraisemblance

- Hypothèses
  - Le processus de substitution suit un modèle probabiliste dont on connaît l'expression mathématique, mais pas les valeurs numériques
  - Les sites évoluent indépendamment les uns des autres
  - Les sites évoluent suivant le même processus = hypothèse d'uniformité (hypothèse pouvant être levée par l'inclusion d'une loi gamma)
  - Le taux de substitution ne change pas au cours du temps, cad le long d'une branche = hypothèse d'homogénéité, mais il peut varier entre les branches
    - En fait, on suppose que le modèle et ses paramètres qualitatifs sont conservés d'une branche à l'autre, mais que seul la quantité d'évolution varie d'une branche à l'autre

Hétérogénéité des vitesses d'évolution

- Sous l'hypothèse que les taux d'évolution sont des variables aléatoires tirées d'une distribution continue, alors on peut la modéliser par une distribution Gamma facilement intégrable par le maximum de vraisemblance.
- Soit, la fonction de densité de la distribution Gamma $g(r) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} r^{\alpha-1} e^{-r/\beta}$
- Le paramètre de forme α va être estimé au maximum de vraisemblance et traité comme un paramètre du vecteur de paramètres Θ

$$L^{(i)} = \int_0^{\infty} g(r) P(D^{(i)} | r, \Theta) dr$$

- Discrétisation de la distribution en K classes

$$L^{(i)} = \frac{1}{K} \sum_{k=1}^{K} P(D^{(i)} | r = r_k, \Theta)$$

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Définition de la vraisemblance

- Données
  - Séquences d'ADN lignées (*n* sites)
  - Modèle d'évolution (JC, K2P, HKY…)
- Hypothèses
  - Soit Θ le vecteur de paramètres: les paramètres du modèle θ , topologie T , Γ, longueurs de branches ℓ
- Décomposition de la vraisemblance
  - $L = P(D|H)$

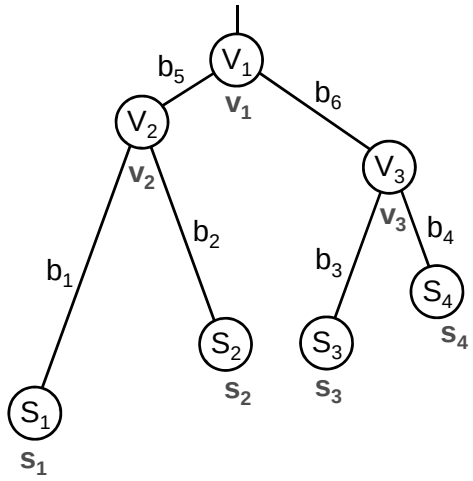
$$L = P(D^{(1)}|H) \times P(D^{(2)}|H) \times ... \times P(D^{(n)}|H)$$

$$L = \Pi_{i=1}^{n} \left[P(D^{(i)} | T, \theta, \ell) \right] = \Pi_{i=1}^{n} \left[P(D^{(i)} | \Theta) \right]$$

La vraisemblance est calculée de manière indépendante à chaque site

Vraisemblance à un site

- Soit s_1, s_2, s_3 et s_4 les nucléotides observés à un site de l'alignement
- Soit v_1, v_2 et v_3 les nucléotides ancestraux présents aux nœuds v_1, v_2 et v_3 => ces états sont inconnus



$$L^{(i)} = \left[P(D^{(i)}|\Theta)\right]$$

$$= P(s_1, s_2, s_3, s_4, v_1, v_2, v_3|\Theta)$$

$$= P(v_1)P(v_2|v_1, b_5)P(s_1|v_2, b_1)P(s_2|v_2, b_2)P(v_3|v_1, b_6)P(s_3|v_3, b_3)P(s_4|v_3, b_4)$$

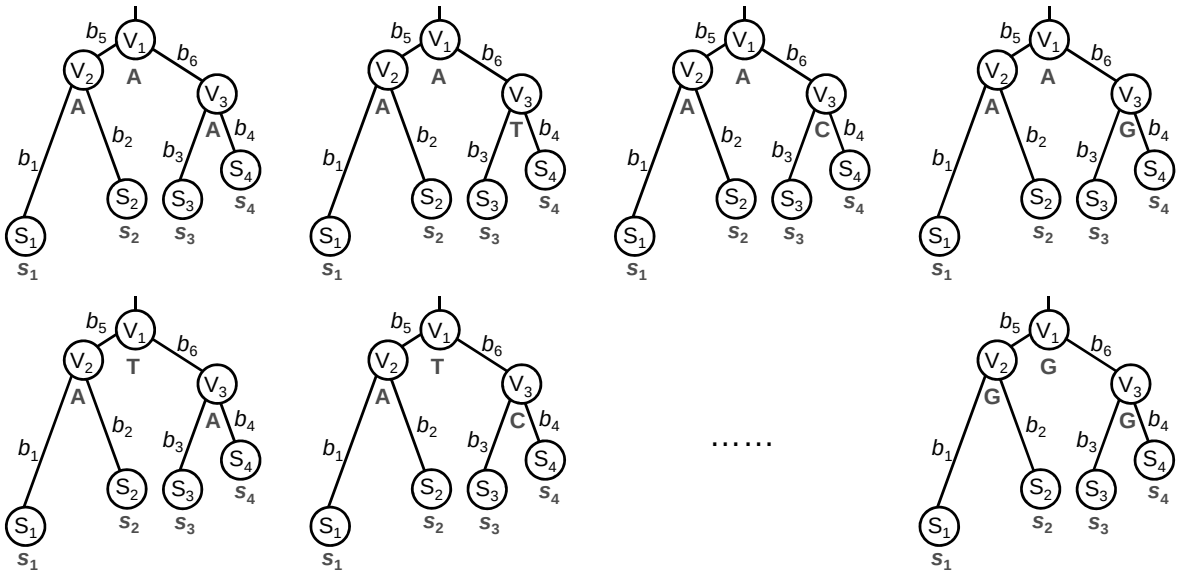
$$= \sum_{v_1}\sum_{v_2}\sum_{v_3} P(s_1, s_2, s_3, s_4, v_1, v_2, v_3|\Theta)$$

$$= \sum_{v_1}\sum_{v_2}\sum_{v_3} P(v_1)P(v_2|v_1, b_5)P(s_1|v_2, b_1)P(s_2|v_2, b_2)P(v_3|v_1, b_6)P(s_3|v_3, b_3)P(s_4|v_3, b_4)$$

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Il y a 64 scénarios possibles à l'origine de s_1, s_2, s_3 et s_4



(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Calcul de la fonction de vraisemblance

- Le calcul de $L^{(i)}$ est réalisé pour chaque site
- Le calcul des probabilités conditionnelles $P(x|y,b)$ nécessite un modèle probabiliste du processus de substitution
- Sous l'hypothèse que les séquences sont à l'équilibre

$$P(v_1) = \pi_{v1}$$

avec π_{v1} la fréquence à l'équilibre de l'état de caractère v_1

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Calcul de la vraisemblance à un site

- Lorsque le nombre de séquences considéré = 4, il y a 64 combinaisons d'états de caractères possibles aux 3 nœuds internes

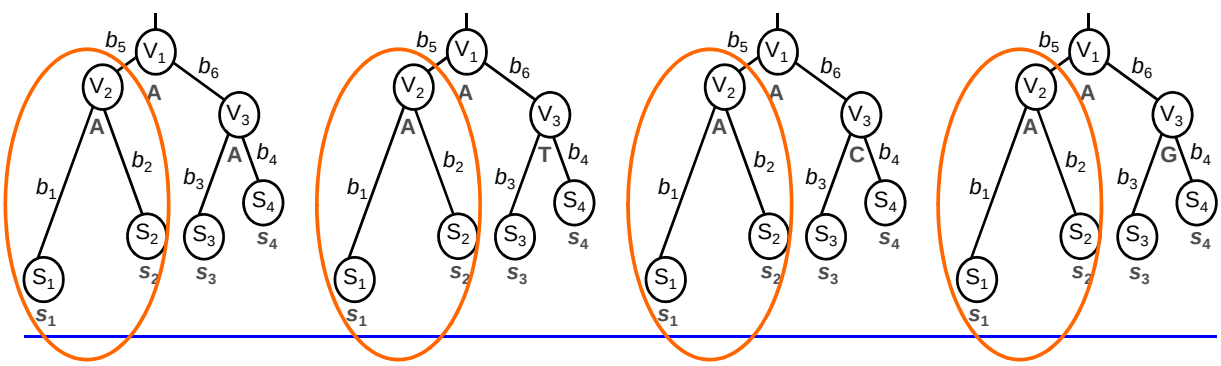
=> Augmente de manière exponentielle lorsque le nombre de séquences augmente

Ex. si n =10 => 4^9 = 262144 combinaisons

si n=30 => 4_{29} ~ 3 x 10^{17} combinaisons

⇒ devient incalculable quand n augmente

⇒ mais … implique le recalcul des mêmes valeurs de nombreuses fois



Algorithme d'élagage

- Felsenstein propose une méthode de calcul appelée **élagage** (pruning)

$$L^{(i)} = \sum_{v1}\sum_{v2}\sum_{v3} P(v1)P(v2|v1,b5)P(s1|v2,b1)P(s2|v2,b2)P(v3|v1,b6)P(s3|v3,b3)P(s4|v3,b4)$$

$$= \sum_{v1} P(v1)\times\left[\sum_{v2} P(v2|v1,b5)P(s1|v2,b1)P(s2|v2,b2)\right]\times\left[\sum_{v3} P(v3|v1,b6)P(s3|v3,b3)P(s4|v3,b4)\right]$$

- Basé sur le calcul de vraisemblances conditionnelles $L_K^{(i)}(k)$ (aussi appelées vraisemblances partielles) à chaque nœud *K* de l'arbre

- Elle correspond à la probabilité d'observer les données aux feuilles du sous arbre raciné par K, sachant l'état de caractère *k* à ce nœud.

Calcul des vraisemblances partielles

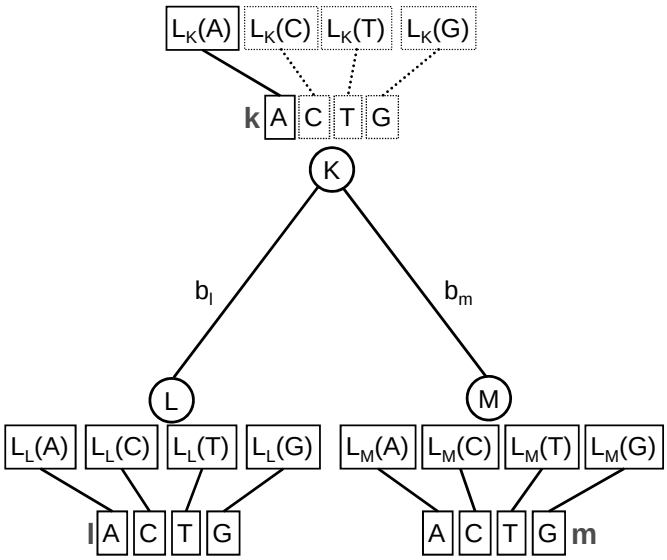
- Si K correspond à une feuille, alors:
  - $L_K^{(i)}(k) = 1$ pour l'un des 4 états de caractères
  - $L_K^{(i)}(k) = 0$ pour les 3 autres états de caractères
- Si K correspond à un nœud, alors:

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

avec *L* et *M* les deux nœuds fils de *K*

- Si K correspond à la racine, alors:

$$L^{(i)} = \sum_{v1} P(v1)L_{v1}^{(i)}(v1)$$

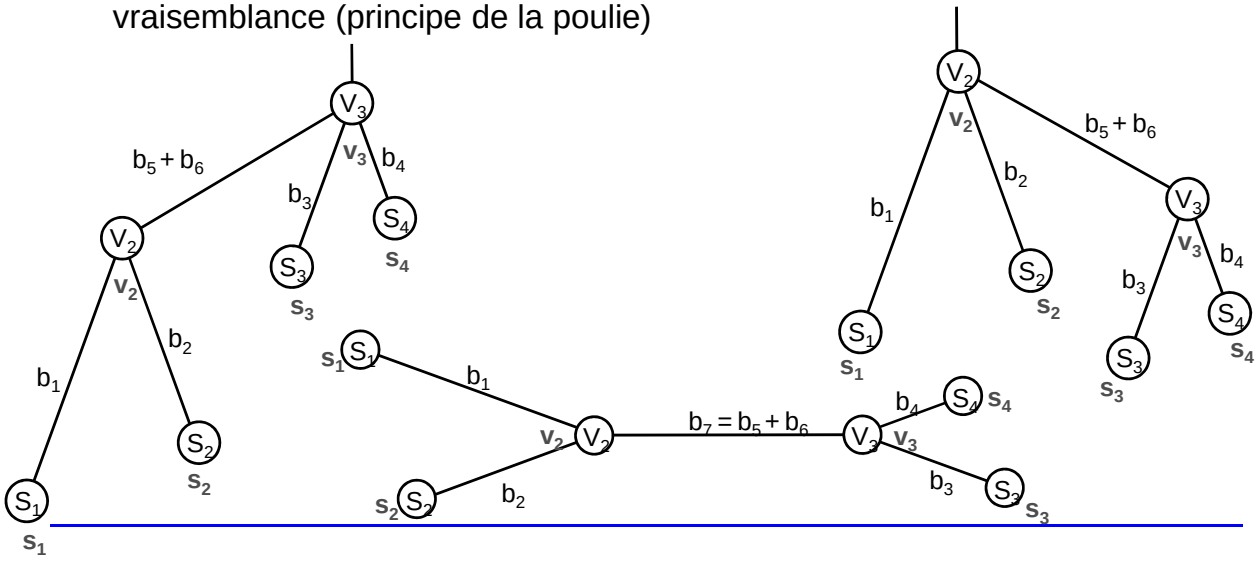


Le processus est réitéré des feuilles jusqu'à la racine

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Influence de la position de la racine

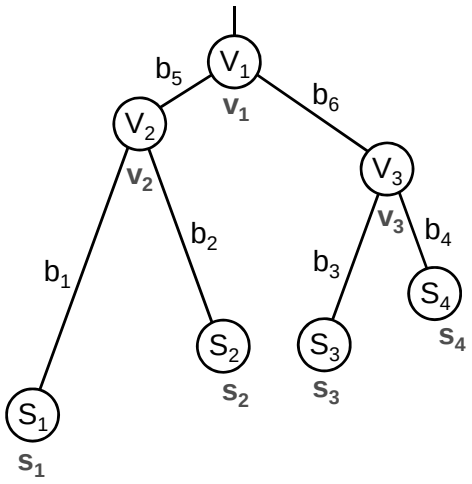
- L'application de l'algorithme d'élagage requière une topologie racinée
- Si le modèle d'évolution est réversible et que les séquences étudiées sont à l' équilibre, alors la position de la racine n'influence pas le calcul de la vraisemblance (principe de la poulie)



Les arbres inférés au maximum de vraisemblance sont non racinés

Application numérique: calcul de la vraisemblance d'un arbre

- Calcul de la vraisemblance de l'arbre ci-contre
- Vecteur de paramètres Θ:
  - Longueurs de branches $b_1 = 0.5$; $b_2 = 0.4$; $b_3 = b_6 = 0.3$; $b_4 = b_5 = 0.2$
  - Modèle d'évolution JC69, avec P(t) la matrice de substitution
  - T la topologie de l'arbre
- Alignement de séquences nucléiques
- Calcul de la vraisemblance à chaque site de l'alignement



Application numérique : Détermination des matrices de substitutions *P(t)*

- Sous le modèle de Jukes et Cantor $p(b) = \frac{3}{4}(1 - e^{-4b/3})$
- Donc

  $p(b_1) = p(0.5) = 0.36$; $p(b_2) = p(0.4) = 0.31$; $p(b_3) = p(b_6) = p(0.3) = 0.25$

  $p(b_4) = p(b_5) = p(0.2) = 0.18$
- Sous le modèle JC69, toutes les substitutions se produisent à la même fréquence $p_{A,T} = P_{A,C} = p_{A,G} = p_1/3$

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

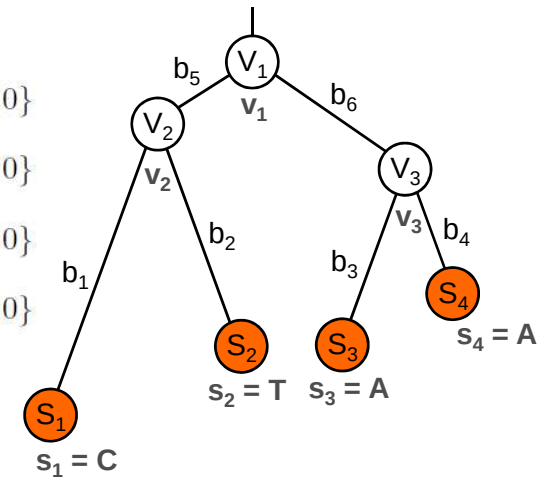
$$P(0.4) = \begin{bmatrix} 0.69 & 0.10 & 0.10 & 0.10 \\ 0.10 & 0.69 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.69 & 0.10 \\ 0.10 & 0.10 & 0.10 & 0.69 \end{bmatrix} \quad P(0.5) = \begin{bmatrix} 0.64 & 0.12 & 0.12 & 0.12 \\ 0.12 & 0.64 & 0.12 & 0.12 \\ 0.12 & 0.12 & 0.64 & 0.12 \\ 0.12 & 0.12 & 0.12 & 0.64 \end{bmatrix}$$

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Application numérique : Calcul des vraisemblances partielles aux feuilles

- Si le site considéré présente les états de caractères S_1 = C ; S_2 = T ; S_3 = S_4 = A, alors les vraisemblances partielles aux feuilles sont égales à

$$\left\{L_{S_1}^{(i)}(\mathrm{A}), L_{S_1}^{(i)}(\mathrm{C}), L_{S_1}^{(i)}(\mathrm{T}), L_{S_1}^{(i)}(\mathrm{G})\right\} = \{0,1,0,0\}$$
$$\left\{L_{S_2}^{(i)}(\mathrm{A}), L_{S_2}^{(i)}(\mathrm{C}), L_{S_2}^{(i)}(\mathrm{T}), L_{S_2}^{(i)}(\mathrm{G})\right\} = \{0,0,1,0\}$$
$$\left\{L_{S_3}^{(i)}(\mathrm{A}), L_{S_3}^{(i)}(\mathrm{C}), L_{S_3}^{(i)}(\mathrm{T}), L_{S_3}^{(i)}(\mathrm{G})\right\} = \{1,0,0,0\}$$
$$\left\{L_{S_4}^{(i)}(\mathrm{A}), L_{S_4}^{(i)}(\mathrm{C}), L_{S_4}^{(i)}(\mathrm{T}), L_{S_4}^{(i)}(\mathrm{G})\right\} = \{1,0,0,0\}$$



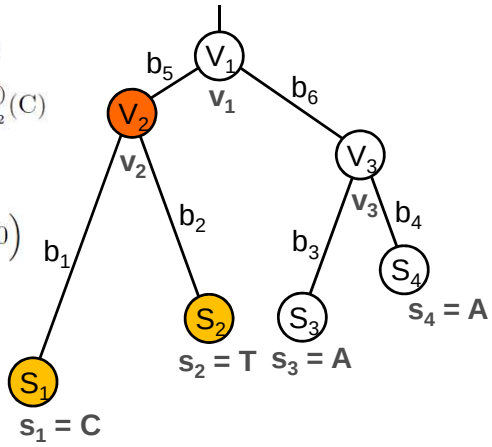
(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Application numérique : Calcul des vraisemblances partielles aux nœuds

- Si le nœud V_2 est considéré, il a pour fils les feuilles S_1 et S_2.

$$L_{V_2}^{(i)}(\mathrm{A}) = \Big[p_{AA}(0.5)L_{S_1}^{(i)}(\mathrm{A}) + p_{AC}(0.5)L_{S_1}^{(i)}(\mathrm{C}) + p_{AT}(0.5)L_{S_1}^{(i)}(\mathrm{T}) + p_{AG}(0.5)L_{S_1}^{(i)}(\mathrm{G})\Big] \times \Big[p_{AA}(0.4)L_{S_2}^{(i)}(\mathrm{A}) + p_{AC}(0.4)L_{S_2}^{(i)}(\mathrm{C}) + p_{AT}(0.4)L_{S_2}^{(i)}(\mathrm{T}) + p_{AG}(0.4)L_{S_2}^{(i)}(\mathrm{G})\Big]$$

$$L_{V_2}^{(i)}(\mathrm{A}) = \Big(0 + p_{AC}L_{S_2}^{(i)}(\mathrm{C}) + 0 + 0\Big) \times \Big(0 + 0 + p_{AT}L_{S_2}^{(i)}(\mathrm{T}) + 0\Big) = 0.12 \times 1 \times 0.10 \times 1 = 0.012$$



- Similairement

$$L_{V_2}^{(i)}(C) = 0.064$$

$$L_{V_2}^{(i)}(T) = 0.083$$

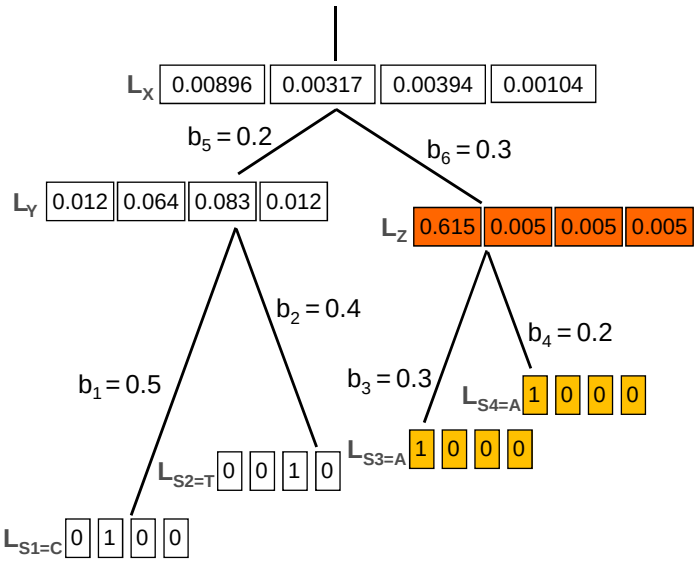
$$L_{V_2}^{(i)}(G) = 0.012$$

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Application numérique : Calcul des vraisemblances partielles aux nœuds

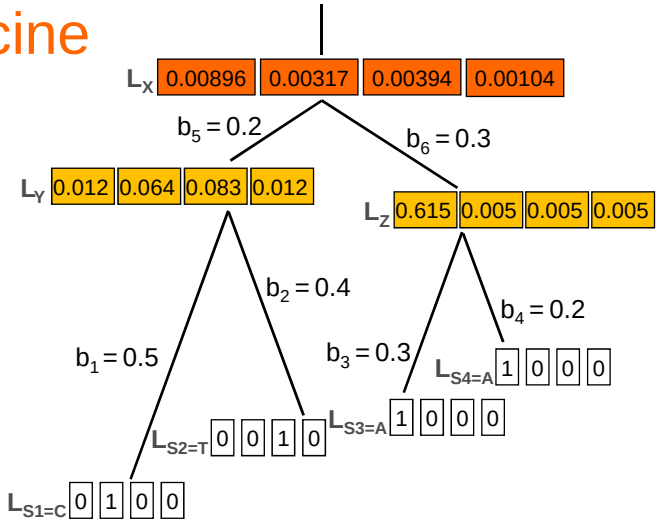
- Application à V_3



(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Application numérique : Calcul de la vraisemblance à la racine

- Calcul de la vraisemblance à la racine



$$L^{(i)} = \sum_{v1} \pi_{v1} L_{V_1}^{(i)}(v_1)$$

$$= \pi_A L_{V_1}^{(i)}(A) + \pi_C L_{V_1}^{(i)}(C) + \pi_T L_{V_1}^{(i)}(T) + \pi_G L_{V_1}^{(i)}(G)$$

$$= 0.25 \times (0.00896 + 0.00317 + 0.00394 + 0.00104)$$

$$= 0.0042775$$

Et donc $\ln(0.0042775) = -5.4544$

(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Application numérique : Calcul de la vraisemblance d'un arbre

- La procédure est répétée pour tous les sites de l'alignement

Maximisation de la vraisemblance

(1) On considère une topologie T , un site et un ensemble de longueurs de branches b (les longueurs de branches initiales sont en générales estimées par un arbre reconstruit par une méthode de distance, ex. NJ)

(2) On calcule la vraisemblance des paramètres = probabilité d'observer les états de caractères au site en fonction des paramètres

(3) On fait le calcul pour tous les caractères (T, θ, b)

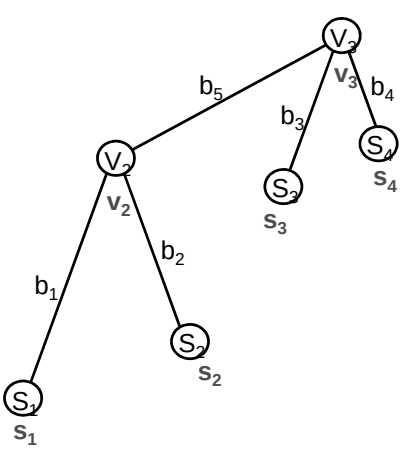
(4) On calcule les longueurs de branches b et les paramètres θ du modèle qui maximisent la vraisemblance (procédure complexe)

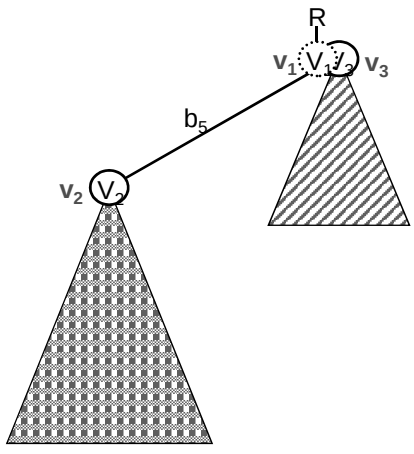
(5) On calcule la vraisemblance pour toutes les topologies possibles

(6) On retient la topologie qui a la plus grande vraisemblance

Maximisation de la vraisemblance (suite)

- En pratique il n'est pas possible de tester toutes les topologies => exploration de l'espace des arbres via les heuristiques vues précédemment.
- Il n'est pas possible de tester toutes les longueurs de branches => optimisation branche par branche

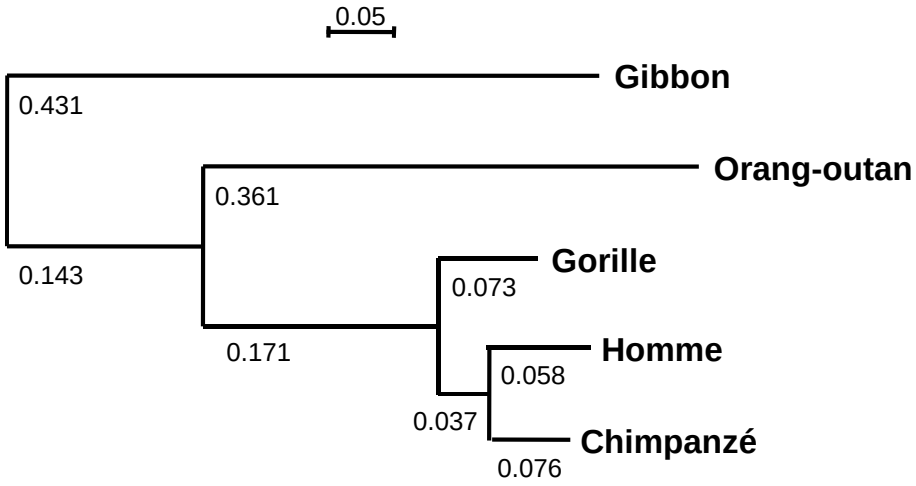




(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Application à la phylogénie des Hominoïdes

- Sous le modèle HKY85 + G, l'arbre le plus vraisemblable est



(Perrière & Brochier-Armanet, (2010) Concepts et méthodes en phylogénie moléculaire, Springer)

Propriétés du maximum de vraisemblance

- C'est une des méthodes les plus justifiées d'un point de vue théorique
- Les simulations montrent que cette méthode est supérieure aux autres dans beaucoup de cas. En particulier elle est moins sensible aux artefacts d'attraction des longues branches
- Coûteuse en temps de calcul
- Impossible d'évaluer tous les arbres ⇔ utilisation d'heuristiques ⇔ n'est plus certain d'obtenir l'arbre le plus vraisemblable
- Des tests statistiques dérivés du maximum de vraisemblance permettent d'évaluer si des topologies ayant une vraisemblance moins bonne que la topologie la plus vraisemblable sont significativement différentes

Cause de l'incongruence/problèmes rencontrés en phylogénie moléculaire

- Problèmes d'échantillonnages
  - *Séquences trop courtes* => effets stochastiques
  - *Échantillonnage taxonomique trop réduit*
- Problèmes liés à la divergence des séquences
  - *Séquences pas assez variables*
  - *Séquences trop divergentes => saturation*
  - *Séquences présentant des taux d'évolution hétérogènes* (Attraction des longues branches)

=> Facteurs non exclusifs !

Analyse de séquences et phylogénie moléculaire (Céline Brochier-Armanet 2015-2016)

Lectures conseillées

- « Biologie évolutive » Thomas, Lefèvre, Raymond (2010) de boeck
- « Evolution » Barton, Briggs, Eisen, Goldstein, Patel (2007) Cold Spring Harbor Laboratory Press
- « Concepts et méthodes en phylogénie moléculaire » Perrière & Brochier-Armanet (2010) Springer collection IRIS