Phylogénie moléculaire

Master Bioinformatique de Rouen

Guy Perrière

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

15-17 mai 2018

Plan

Quelques grandes étapes I

- Première phylogénie moléculaire (Doolittle et Blombäck, 1964).
- Hypothèse de l'horloge moléculaire (Zuckerkandl et Pauling, 1965).
- Application de la parcimonie aux séquences (Camin et Sokal, 1965).
- Approximation des moindres carrés (Fitch et Margoliash, 1967).
- Premier modèle d'évolution pour les séquences d'ADN (Jukes et Cantor, 1969).
- Algorithme efficace pour la parcimonie (Fitch, 1971).
- Maximum de vraisemblance appliqué aux séquences (Neyman, 1971).

Quelques grandes étapes II

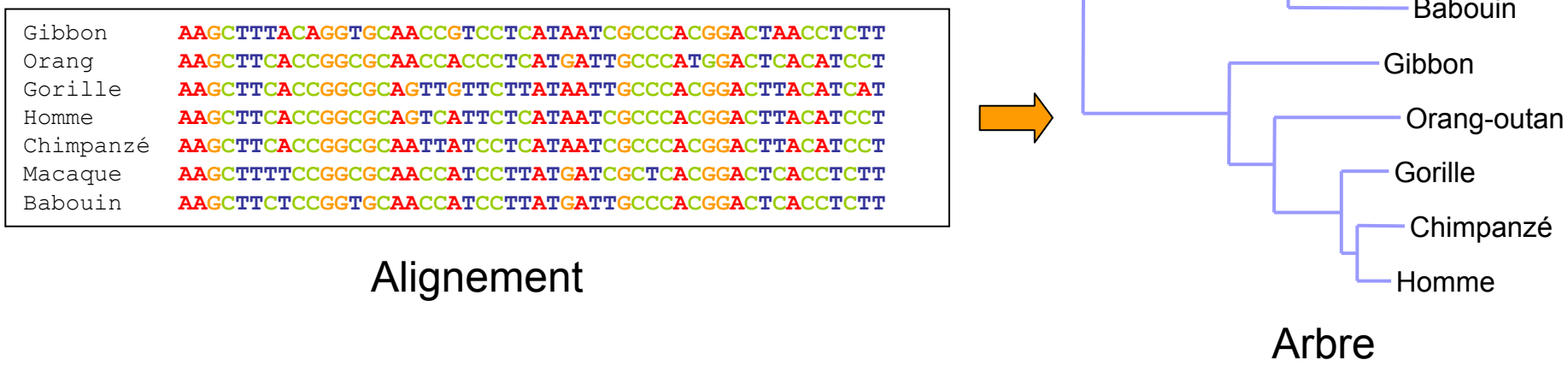
- Modèle PAM pour les séquences d'acides aminés (Dayhoff *et al.*, 1978)
- Modèle de Kimura à deux paramètres (Kimura, 1980).
- Premier algorithme efficace pour le maximum de vraisemblance (Felsenstein, 1981).
- Introduction du *bootstrap* (Felsenstein, 1985).
- Méthode du *Neighbor Joining* (Saitou et Nei, 1987).
- Modélisation de l'hétérogénéité des vitesses d'évolution (Yang, 1994).
- Première phylogénie construite par une approche bayésienne (Yang et Rannala, 1996).

À quoi ça sert ?

- Histoire évolutive de familles de gènes :
  - Analyse des duplications et des pertes de gènes.
  - Détection de transferts horizontaux.
  - Histoire évolutive des organismes les portant.
- Écologie :
  - Phylogéographie.
  - Co-évolution hôte-parasite.
- Épidémiologie.
- Assignation taxonomique ou fonctionnelle.
- Identification de chimères.

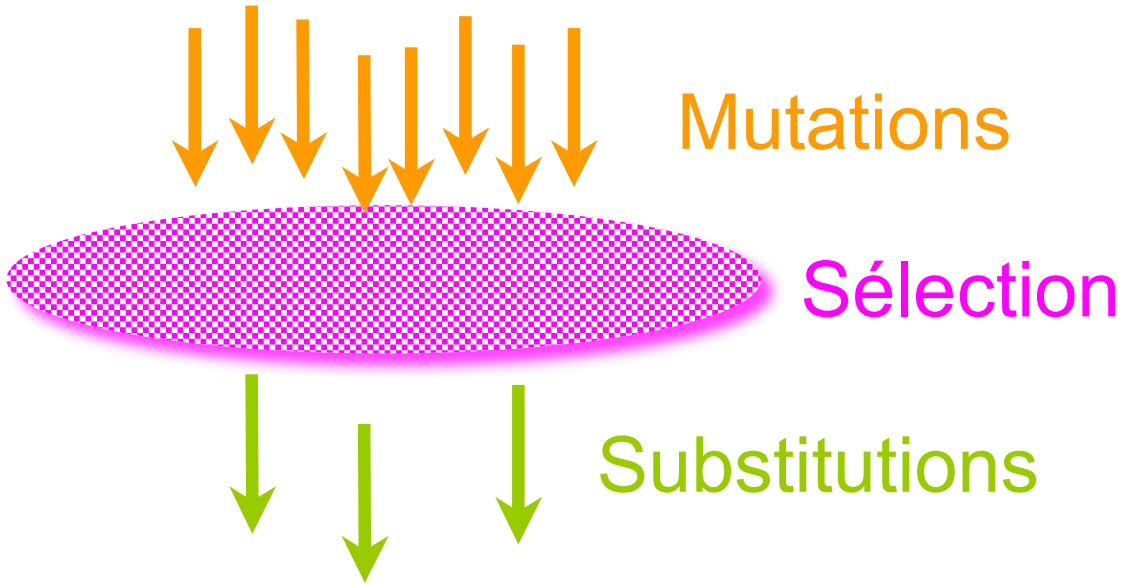
Les données

- Point de départ :
  - Un ensemble de séquences *homologues* alignées.
  - Chaque position dans l'alignement constitue un *site*.
- Résultat obtenu :
  - Un arbre décrivant les relations évolutives entre les séquences (*i.e.*, un arbre phylogénétique).



Mutations et substitutions

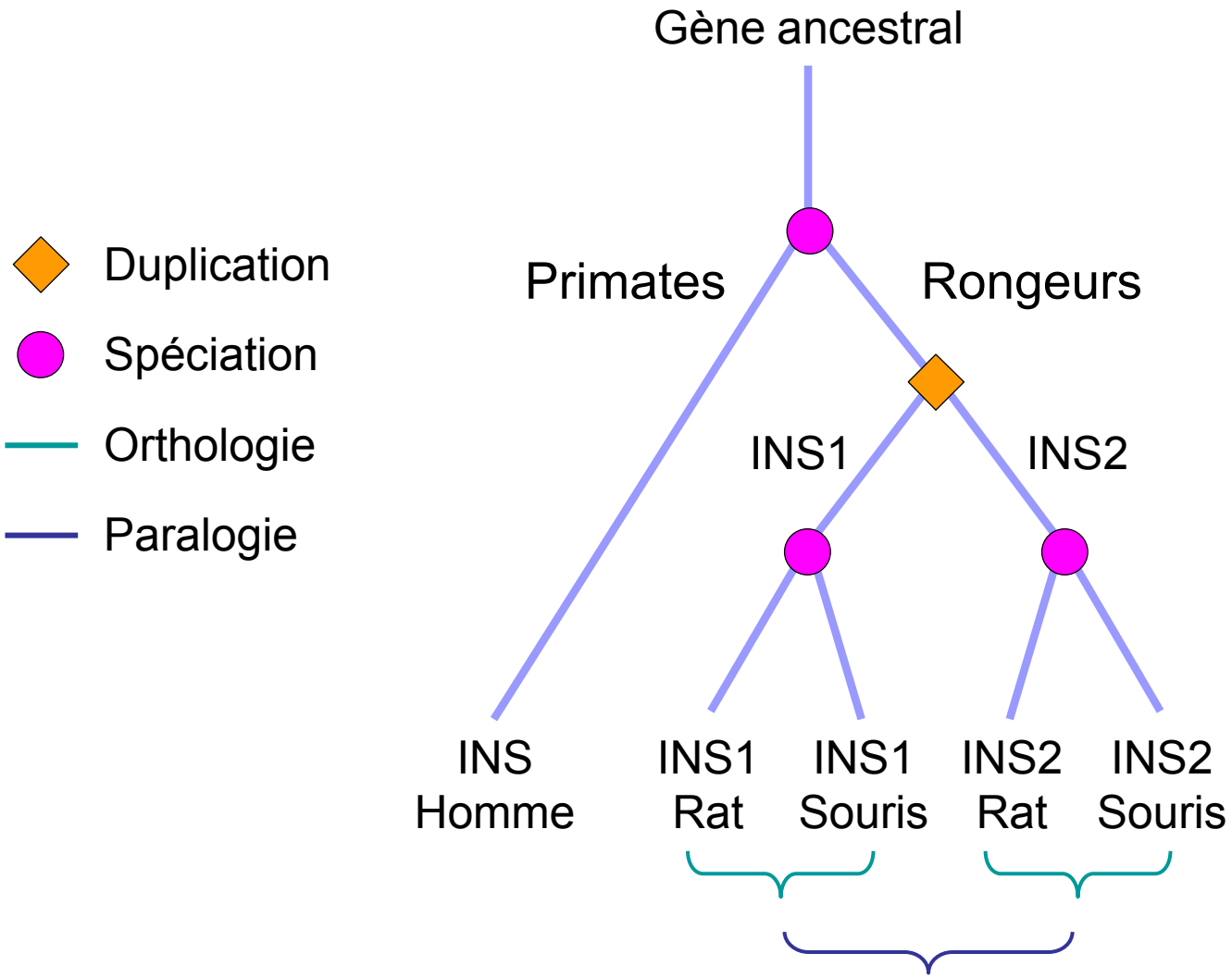
- La grande majorité des *mutations* sont soit neutres (*i.e.*, n'ont aucun effet sur le phénotype), soit délétères :
  - Les mutations avantageuses sont très rares.
- Les *substitutions* correspondent aux mutations qui ont passé le crible de la sélection naturelle.



Homologie ou similarité ?

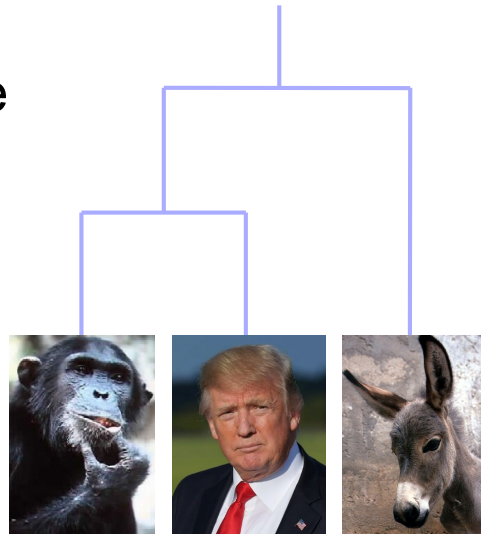
- La phylogénie moléculaire est fondée sur l'utilisation de séquences homologues :
  - Deux séquences sont dites homologues si et seulement si elles possèdent un ancêtre commun.
  - L'existence d'un ancêtre commun est inférée à partir de la similarité.
  - Seuil variable suivant les circonstances :
    - Similarité sans homologie (convergence, répétitions).
    - Homologie avec faible similarité (limitation à quelques positions clés dans les séquences).

Orthologues et paralogues

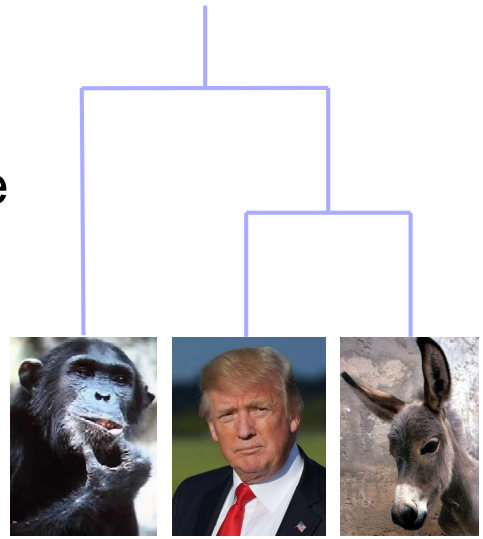


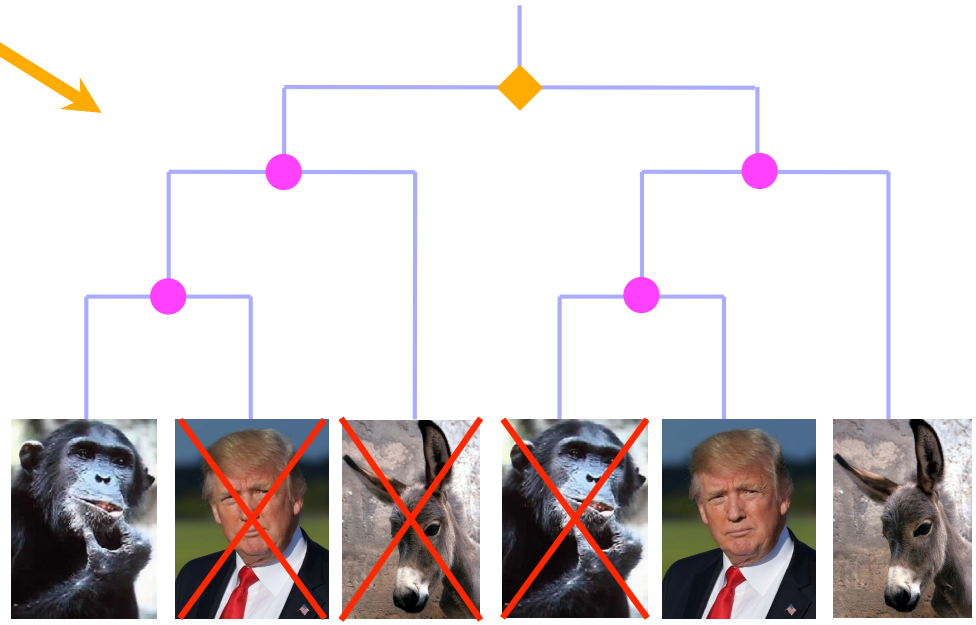
Duplications et phylogénie

Phylogénie vraie



Phylogénie déduite





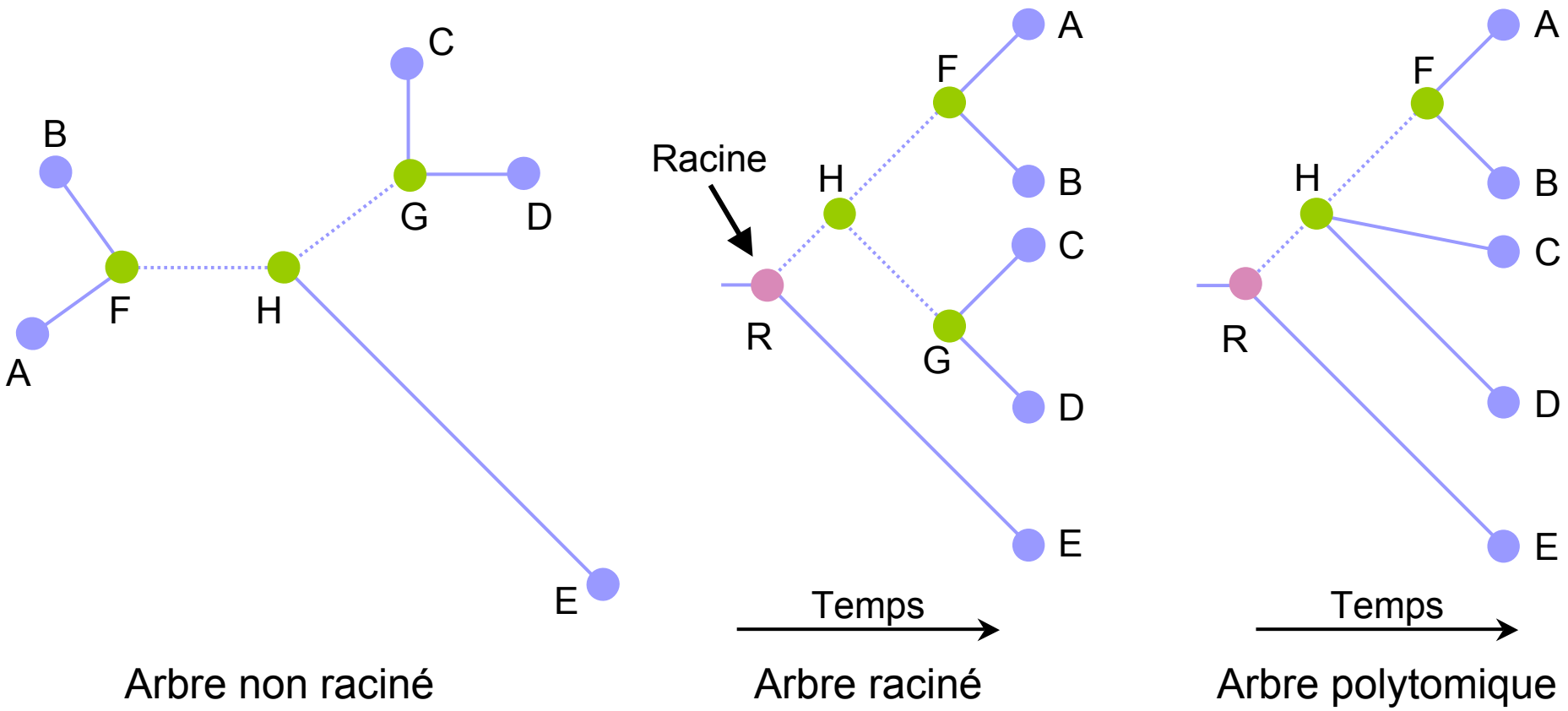
Duplication

Spéciation

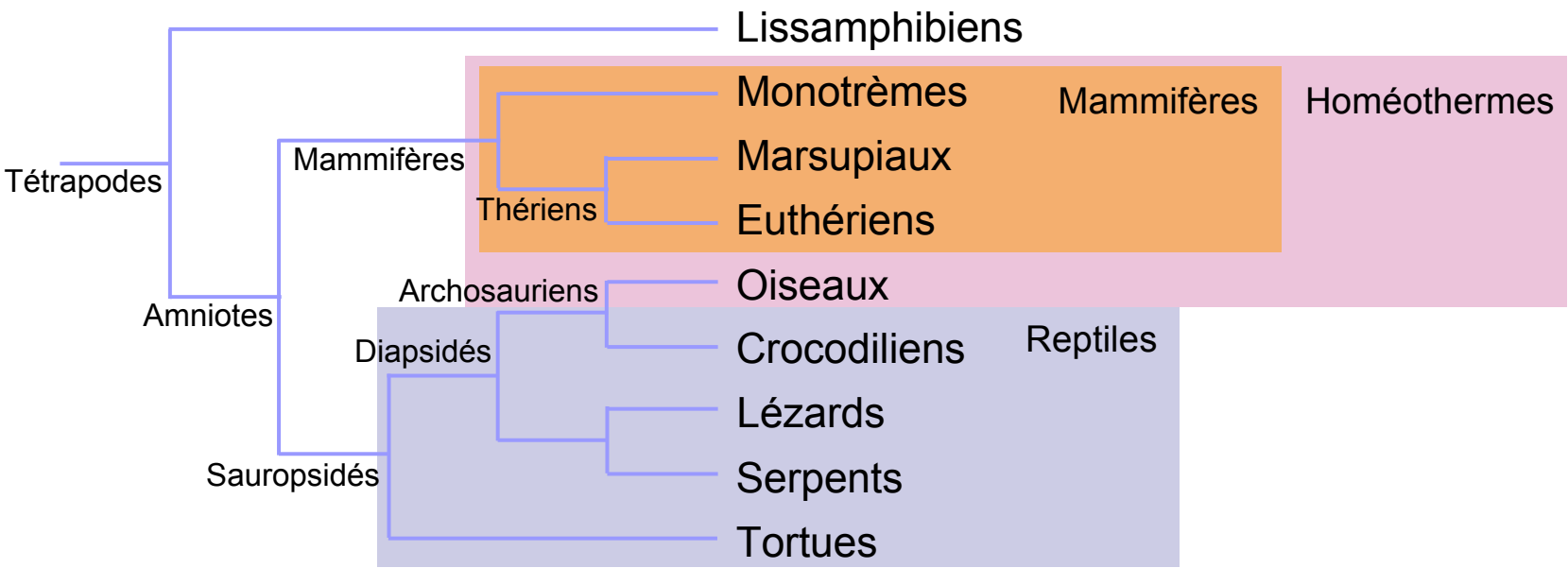
Les paralogues sont fréquents

- Nombre très important même chez les organismes unicellulaires :
  - 30% des gènes d'*E. coli* K12.
  - 40% en moyenne chez les mammifères.
- Existence de duplications multiples :
  - Les relations d'orthologie sont souvent non bijectives.
- Divergences pouvant être importantes après duplication :
  - Difficulté à identifier de nombreux paralogues.

Typologie


A
B
C
D
E
F
G
H
Racine
R
Temps
Arbre non raciné
Arbre raciné
Arbre polytomique
Unité Taxonomique Opérationelle (UTO)
Unité Taxonomique Hypothétique (UTH)

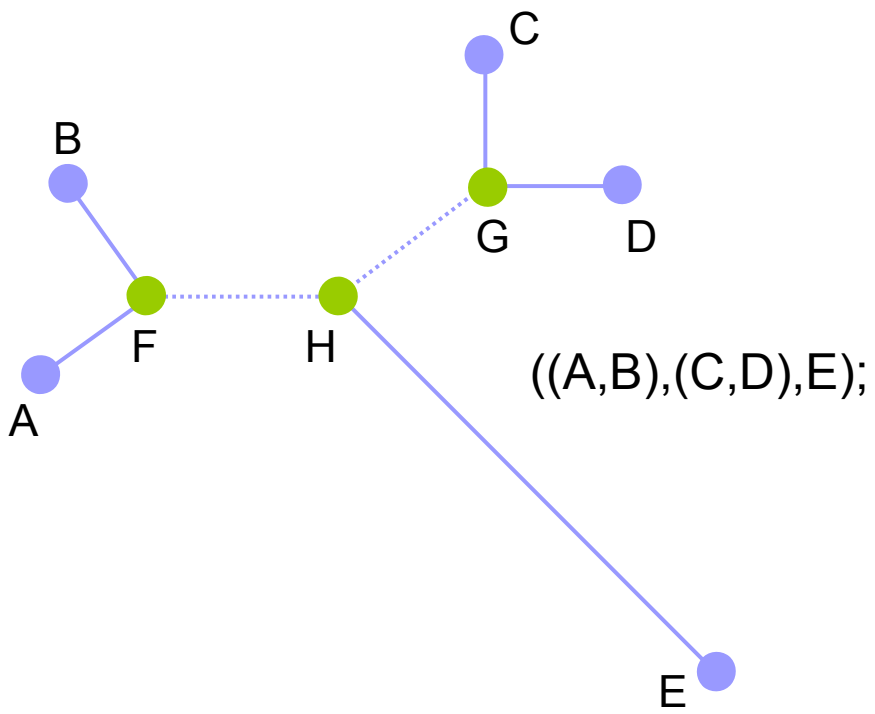
Mono-, poly- et paraphylie



- Dans cette phylogénie des Tétrapodes :
  - Les Mammifères sont *monophylétiques*.
  - Les Homéothermes sont *polyphylétiques*.
  - Les Reptiles (au sens ancien du terme) sont *paraphylétiques*.

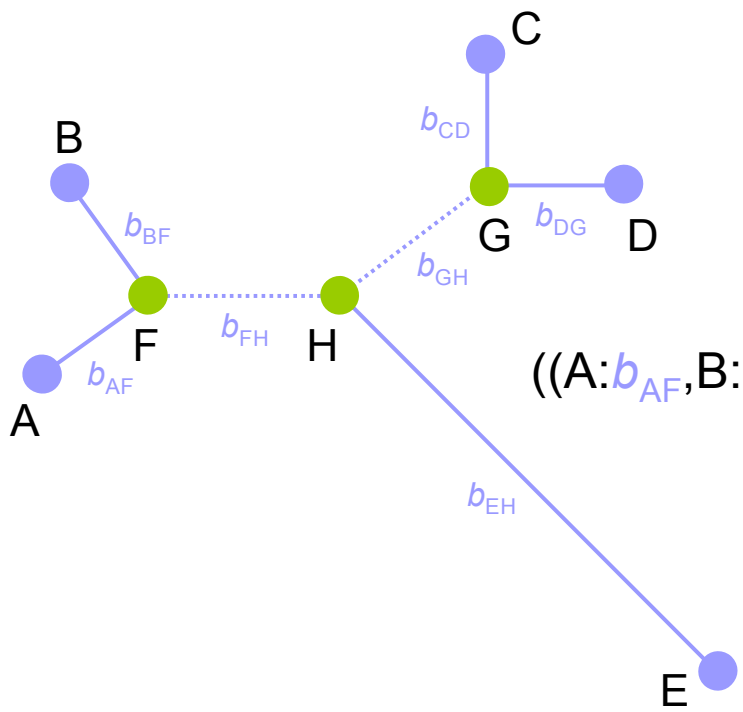
Format Newick standard

- Les UTO (ou groupes d'UTO) descendant d'un même nœud sont placées entre parenthèses.
- Les UTO et groupes d'UTO sont séparés par des virgules.
- La fin de l'arbre est indiquée par un point-virgule.



Extensions courantes

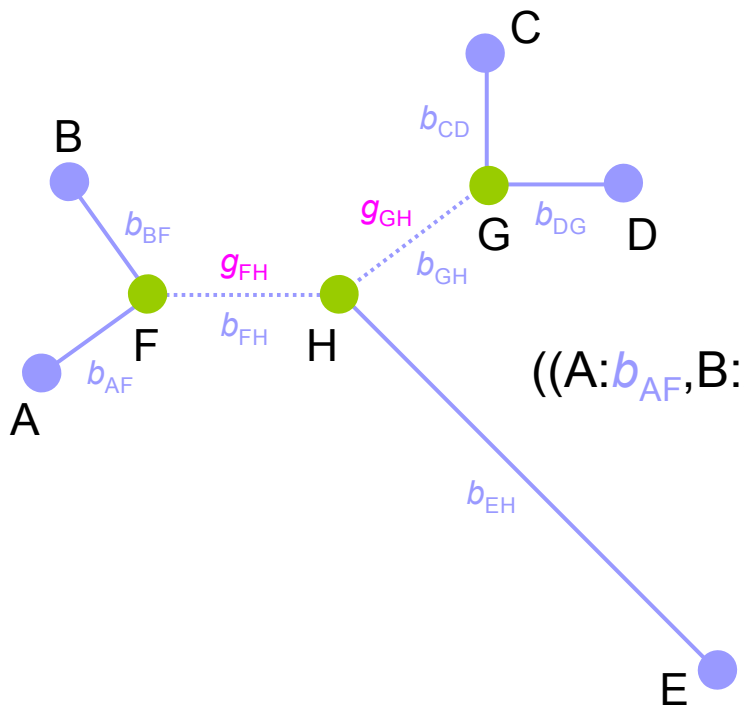
- Longueurs des branches indiquées par leur valeur précédée de deux points.



((A:b_{AF},B:b_{BF}):b_{FH},(C:b_{CG},D:b_{DG}):b_{GH},E:b_{EH});

Extensions courantes

- Longueurs des branches indiquées par leur valeur précédée de deux points.
- Robustesses des branches internes indiquées par un nombre localisé après les parenthèses fermantes délimitant les groupes.



$((A:b_{AF},B:b_{BF})g_{FH}:b_{FH},(C:b_{CG},D:b_{DG})g_{GH}:b_{GH},E:b_{EH});$

Nombre d'arbres racinés

- Soit $B_r^{(n)}$ le nombre d'arbres racinés à n UTO :
- Pour construire un arbre raciné à n UTO, il suffit d'ajouter une UTO à un arbre raciné à $n-1$ UTO.
- Un arbre raciné à $n-1$ UTO possède $n-1$ branches terminales et $n-2$ branches internes, soit $2n-3$ branches au total.
- On en déduit la formule de récurrence :

$$\begin{aligned} B_r^{(n)} &= (2n-3)B_r^{(n-1)} \\ &= (2n-3) \times (2n-5) \times \cdots \times 9 \times 7 \times 5 \times 3 \times 1 \end{aligned}$$

Il est ensuite facile de démontrer que :

$$B_r^{(n)} = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Nombre d'arbres non racinés

- Soit $B_u^{(n)}$ le nombre d'arbres non racinés à n UTO.
- Pour construire un arbre non raciné à n UTO, il suffit d'ajouter une UTO à un arbre non raciné à $n-1$ UTO.
- Un arbre non raciné à $n-1$ UTO possède $n-1$ branches terminales et $n-4$ branches internes, soit $2n-5$ branches au total.
- On en déduit la formule de récurrence :

$$\begin{aligned} B_u^{(n)} &= (2n-5)B_u^{(n-1)} \\ &= (2n-5) \times (2n-7) \times \cdots \times 9 \times 7 \times 5 \times 3 \times 1 \end{aligned}$$

De la même façon que précédemment, on en déduit que :

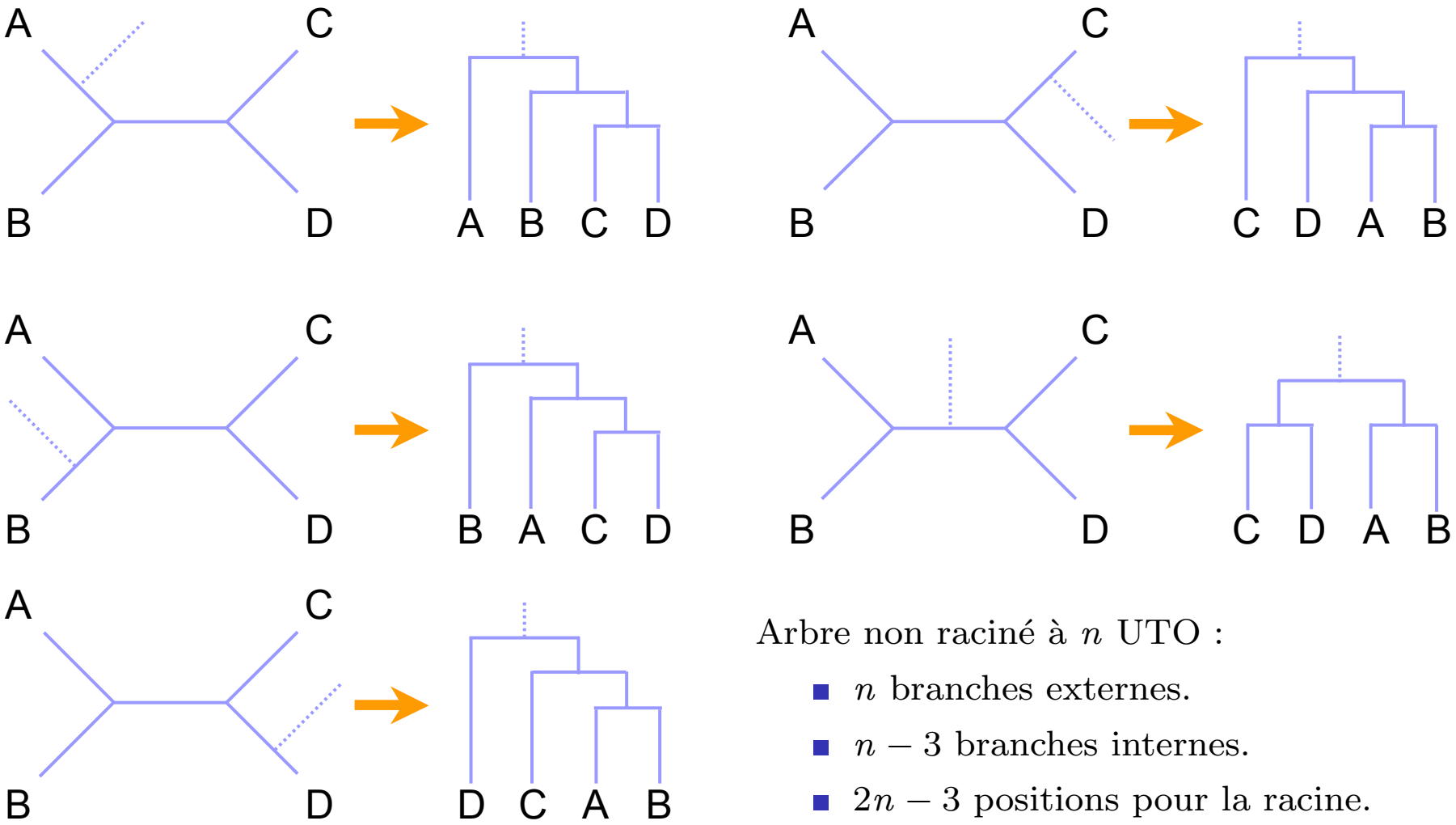
$$B_u^{(n)} = \frac{(2n-5)!}{2^{n-3}(n-3)!} = B_r^{(n-1)}$$

L'arbre caché dans la forêt

- Le nombre d'arbres (racinés ou non) croît donc extrêmement rapidement :
  - Retrouver le bon arbre est pratiquement impossible dès que $n \geq 12$.

n	$B_r^{(n)}$	$B_u^{(n)}$
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025
15	$\approx 2.13 \times 10^{14}$	$\approx 7.91 \times 10^{12}$
20	$\approx 8.20 \times 10^{21}$	$\approx 2.22 \times 10^{20}$
30	$\approx 4.95 \times 10^{38}$	$\approx 8.69 \times 10^{36}$
50	$\approx 2.75 \times 10^{76}$	$\approx 2.84 \times 10^{74}$

Position de la racine



Arbre non raciné à n UTO :

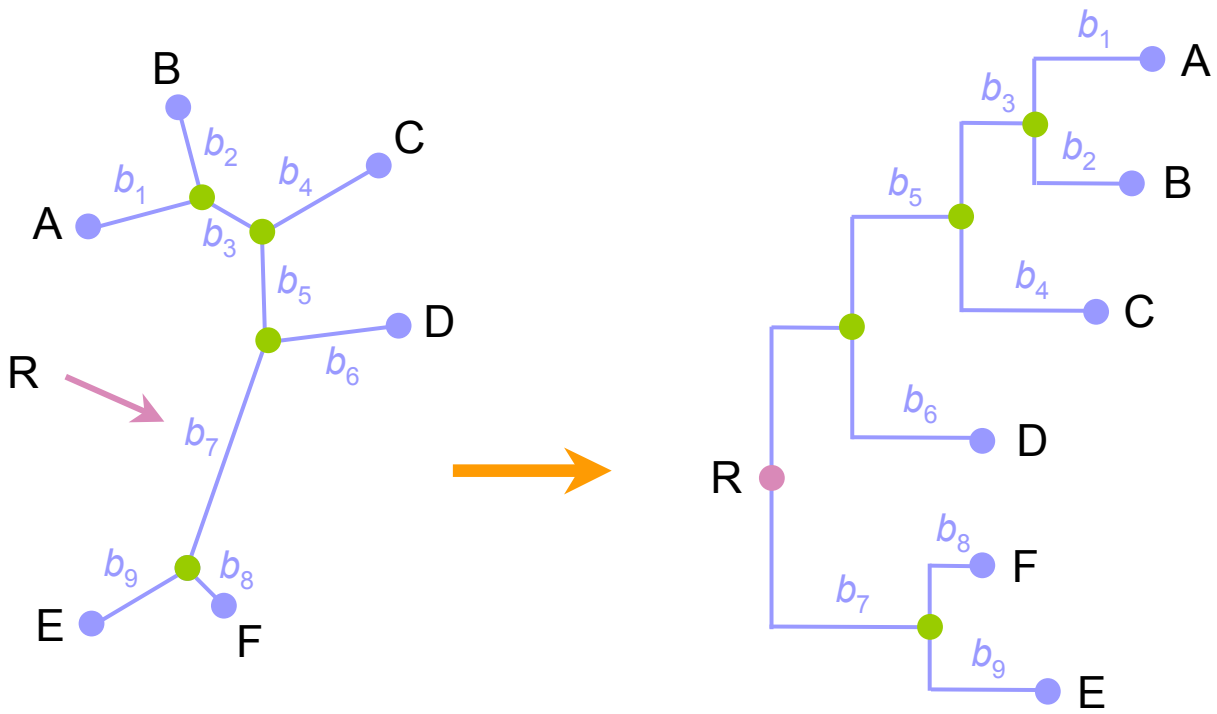
- n branches externes.
- $n - 3$ branches internes.
- $2n - 3$ positions pour la racine.

Racinement d'un arbre

- La plupart des méthodes produisent des arbres sans racine :
  - Pas d'estimation de la direction des changements au cours du temps.
- Plusieurs méthodes de racinement existent :
  - Au point moyen :
    - Hypothèse que toutes les séquences ont évolué à la même vitesse depuis leur divergence avec l'ancêtre commun.
  - À l'aide d'un groupe externe (*outgroup*) fixé *a priori* et connu comme étant extérieur aux taxons étudiés.
  - En utilisant un paralogue.

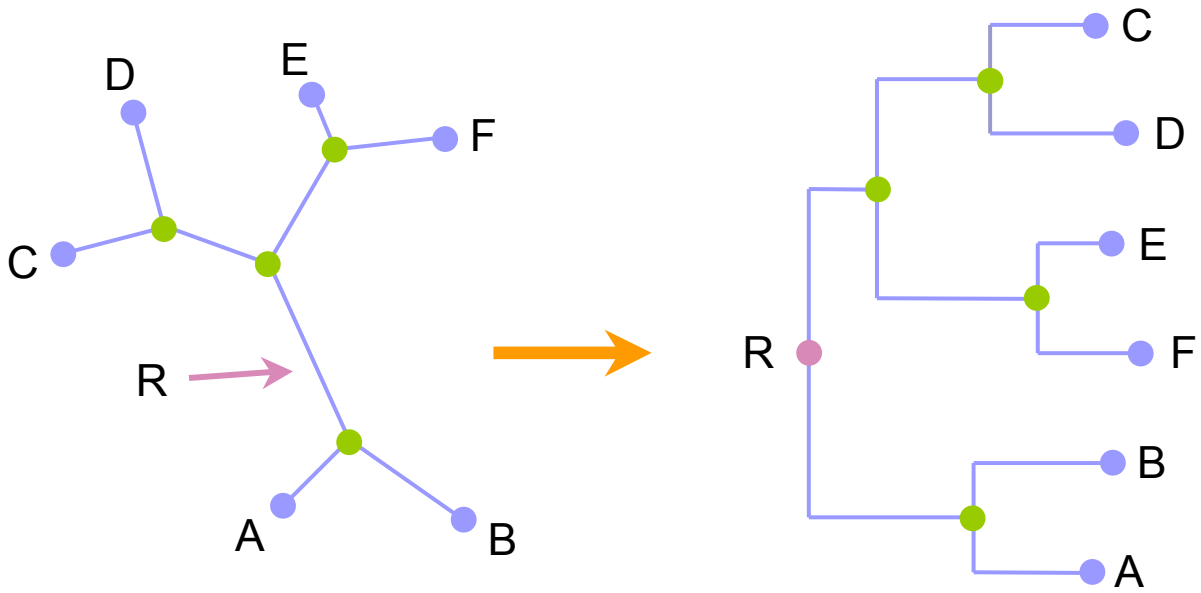
Racinement au point moyen

- Détermination des deux UTO les plus distantes dans l'arbre :
  - Placement de la racine au milieu du chemin.
- Dans l'arbre ci-dessous, A et E sont les deux UTO les plus éloignées et le racinement au point moyen donne :



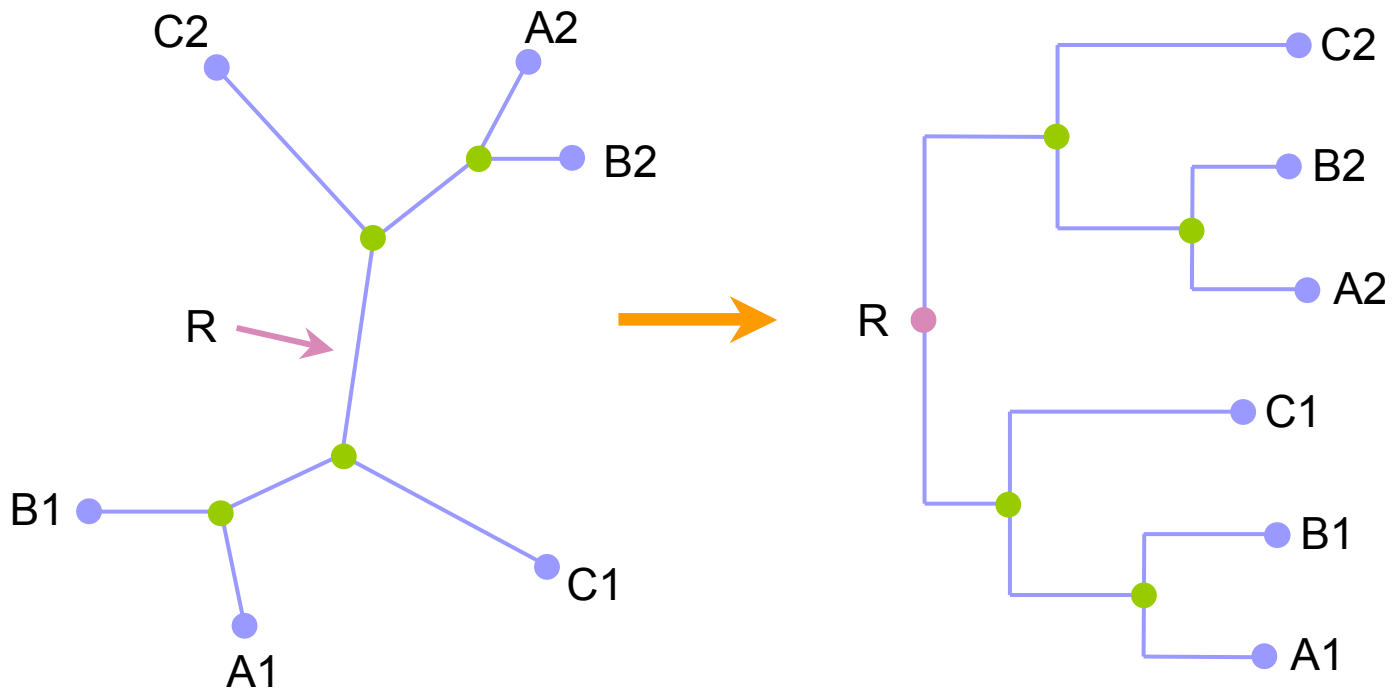
Racinement par un groupe externe

- Choix du groupe externe :
  - Une espèce ou un groupe d'espèces monophylétique qui ne soit ni trop proche ni trop éloigné des organismes d'intérêt.
- Racinement par le groupe {A, B}, supposé extérieur aux organismes d'intérêt que sont C, D, E et F :



Racinement par un paralogue

- Duplication chez l'ancêtre commun à l'ensemble des organismes étudiés :
  - Racinement en utilisant une des deux copies paralogues.
  - Utilisé pour la construction de phylogénies « universelles » (*i.e.*, regroupant les trois domaines du vivant).

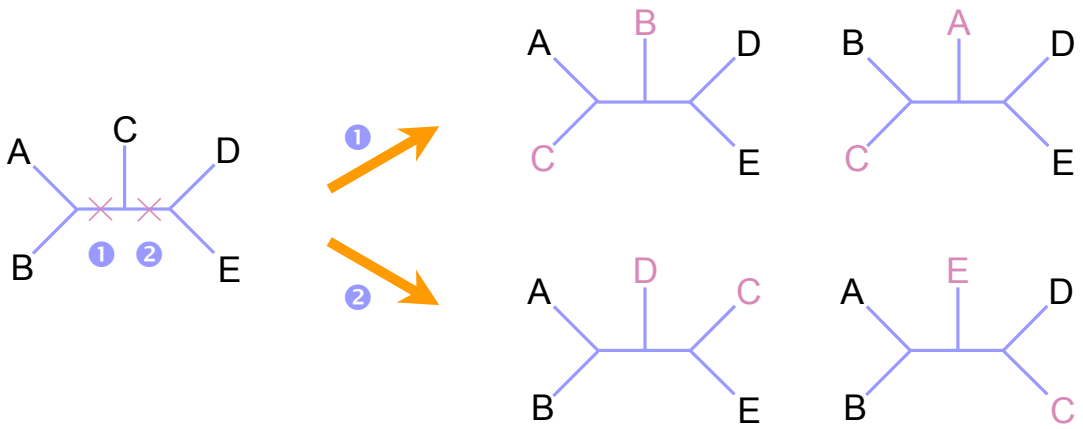


Exploration des topologies

- Plusieurs méthodes de reconstruction phylogénétique nécessitent théoriquement d'évaluer l'ensemble des topologies.
- Différentes approches en fonction du nombre d'UTO :
  - $n < 12$: recherche exhaustive.
  - $n < 20$: algorithme *branch-and-bound*.
  - $n \geq 20$: utilisation d'heuristiques.
- Dans le cas des heuristiques, recherche limitée à une sous-partie de l'ensemble des topologies :
  - Initialisation en utilisant une topologie supposée proche de celle de l'arbre à retrouver.
  - Réarrangements en utilisant différentes méthodes :
    - *Nearest Neighbor Interchange* (NNI).
    - *Subtree Pruning and Regrafting* (SPR).
    - *Tree Bisection and Reconnection* (TBR).

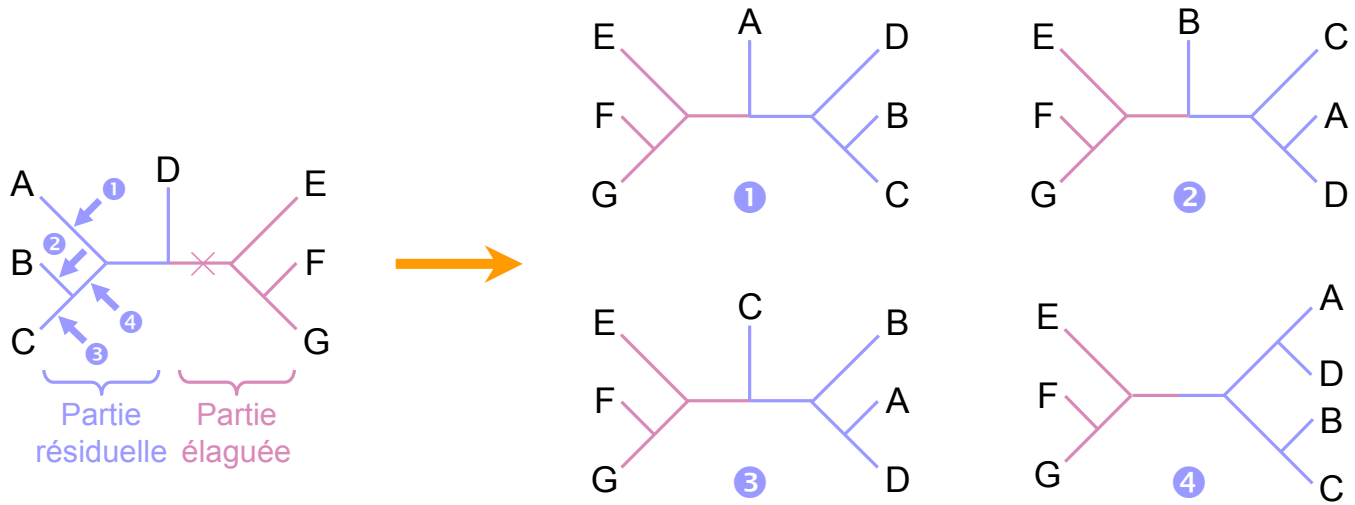
Principe du NNI

- Examen de tous les arbres se situant à une distance topologique $d_T = 2$ de l'arbre de départ :
  - $2n - 6$ réarrangements à effectuer au total.
  - Complexité en $O(n)$.
  - Méthode la plus rapide et la plus répandue.



Principe du SPR

- Coupure de l'arbre de départ au niveau d'une branche interne ou d'une branche externe :
  - Obtention de deux sous-arbres (partie élaguée et partie résiduelle).
  - Placement successif de la partie élaguée sur chacune des branches internes ou externes de la partie résiduelle :



Nombre de réarrangements

- Réitération en échangeant la partie résiduelle et la partie élaguée :
  - Nombre de réarrangements pour une branche interne :

$$(2n_1 - 3 - 1) + (2n_2 - 3 - 1) = 2n - 8$$

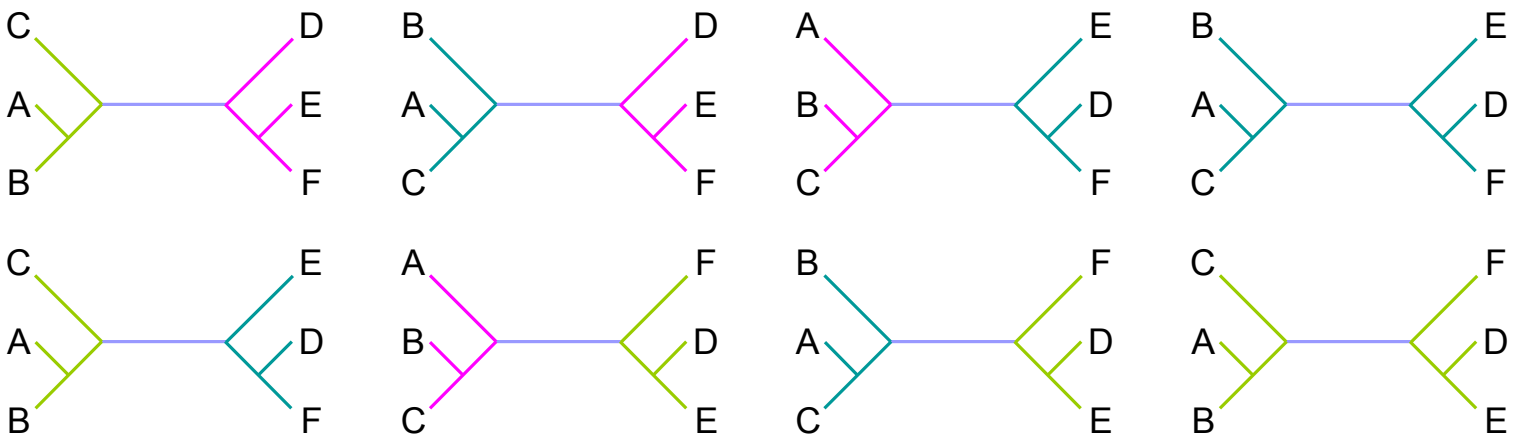
    avec n_1 et n_2 le nombre d'UTO présents de part et d'autre de la branche considérée ($n_1 + n_2 = n$).
  - De la même façon, on démontre que le nombre de réarrangements pour une branche externe est égal à $2n - 6$.
- Réitération du processus complet pour chacune des n branches externes et des $n - 3$ branches internes de l'arbre :
  - Le nombre total de réarrangements est donc égal à :

$$n(2n - 6) + (n - 3)(2n - 8) = 4(n - 3)(n - 2)$$

  - Complexité en $O(n^2)$.

Principe du TBR

- Variante du SPR dans laquelle les deux sous-arbres résultant d'une coupure sont considérés comme étant indépendants :
  - Réalisation de toutes les connexions possibles entre chacune des branches à l'intérieur des deux sous-arbres :



Nombre de réarrangements

- Nombre de réarrangements pour une branche interne :

$$(2n_1 - 3)(2n_2 - 3) - 1$$

avec n_1 et n_2 le nombre d'UTO présents de part et d'autre de la branche considérée.

- Pas de formule générale du calcul du nombre total de réarrangements possibles :
  - Dépendance en fonction de la topologie de départ considérée.
  - Le nombre maximum de réarrangements possibles est égal à :

$$(2n - 3)(n - 3)^2$$

  - Complexité en $O(n^3)$.

Distance de Robinson et Foulds

- Il existe plusieurs mesures de distances topologiques entre deux arbres construits avec des ensembles d'UTO identiques.
- La distance de Robinson et Foulds (1981) est la plus répandue et elle se calcule au moyen de la formule :

$$d_T = 2(b_t - b_c)$$

avec b_t le nombre total de branches internes et b_c le nombre de branches internes présentant des bipartitions identiques entre les deux arbres.

- On en déduit que la distance maximale possible entre deux arbres à n UTO est égale à :

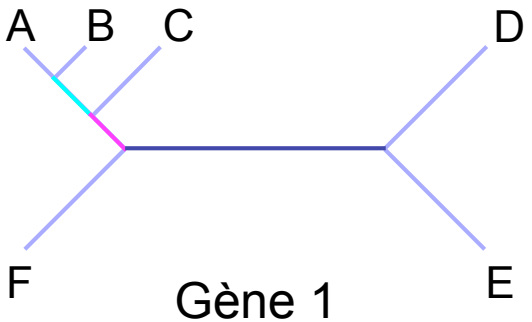
$$d_{T_{\max}} = 2b_t = 2(n-3)$$

Distance normalisée

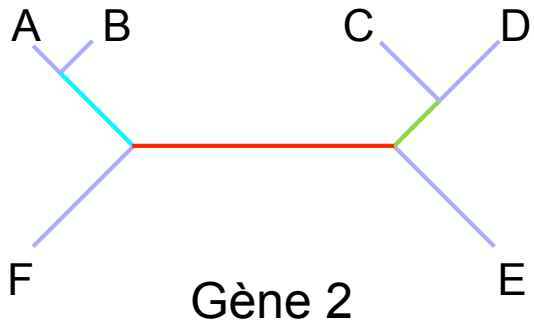
- La distance de Robinson-Foulds standard dépend du nombre de branches internes des arbres étudiés :
  - Pas de comparaisons possibles entre des arbres ne possédant pas le même nombre d'UTO.
- Utilisation d'une valeur normalisée, comprise entre 0 et 1 :

$$\begin{aligned} \tilde{d_T} &= d_T / d_{T_{\max}} \\ &= 2(b_t - b_c)/2b_t \\ &= 1 - b_c/b_t \end{aligned}$$

Exemple



Gène 1



Gène 2

$b_t = 3$ {AB|CDEF} {ABC|DEF} {ABCF|DE}

$b_c = 1$

{AB|CDEF} {ABF|CDE} {ABEF|CD} $b_t = 3$

$$d_T = 2(b_t - b_c) = 4$$

$$\tilde{d}_T = 1 - b_c / b_t = 0.667$$

Plan

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.

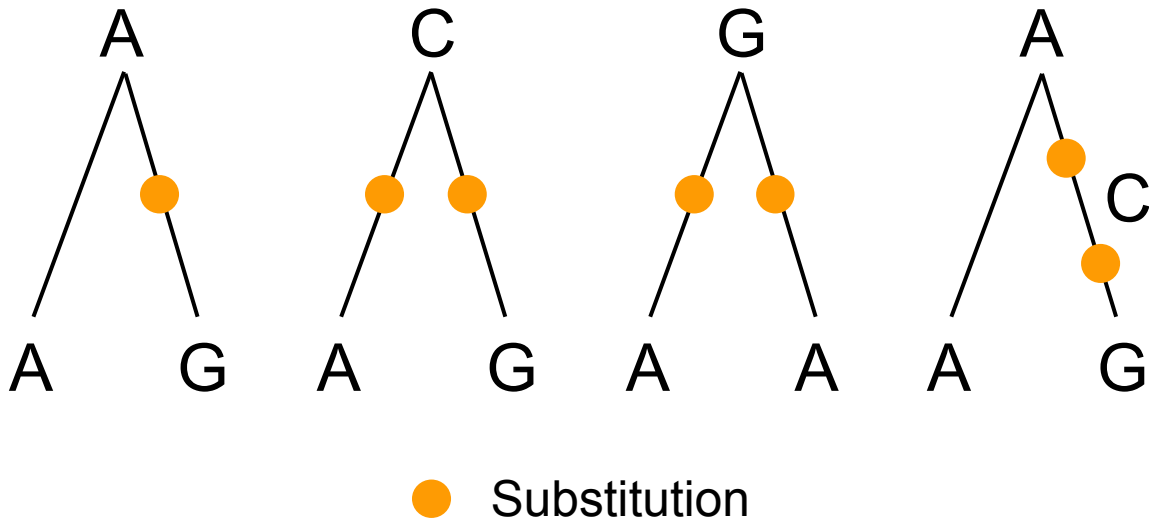
- Variance de l'estimation :

$$\mathbb{V}(p) = \frac{p(1-p)}{\ell}$$

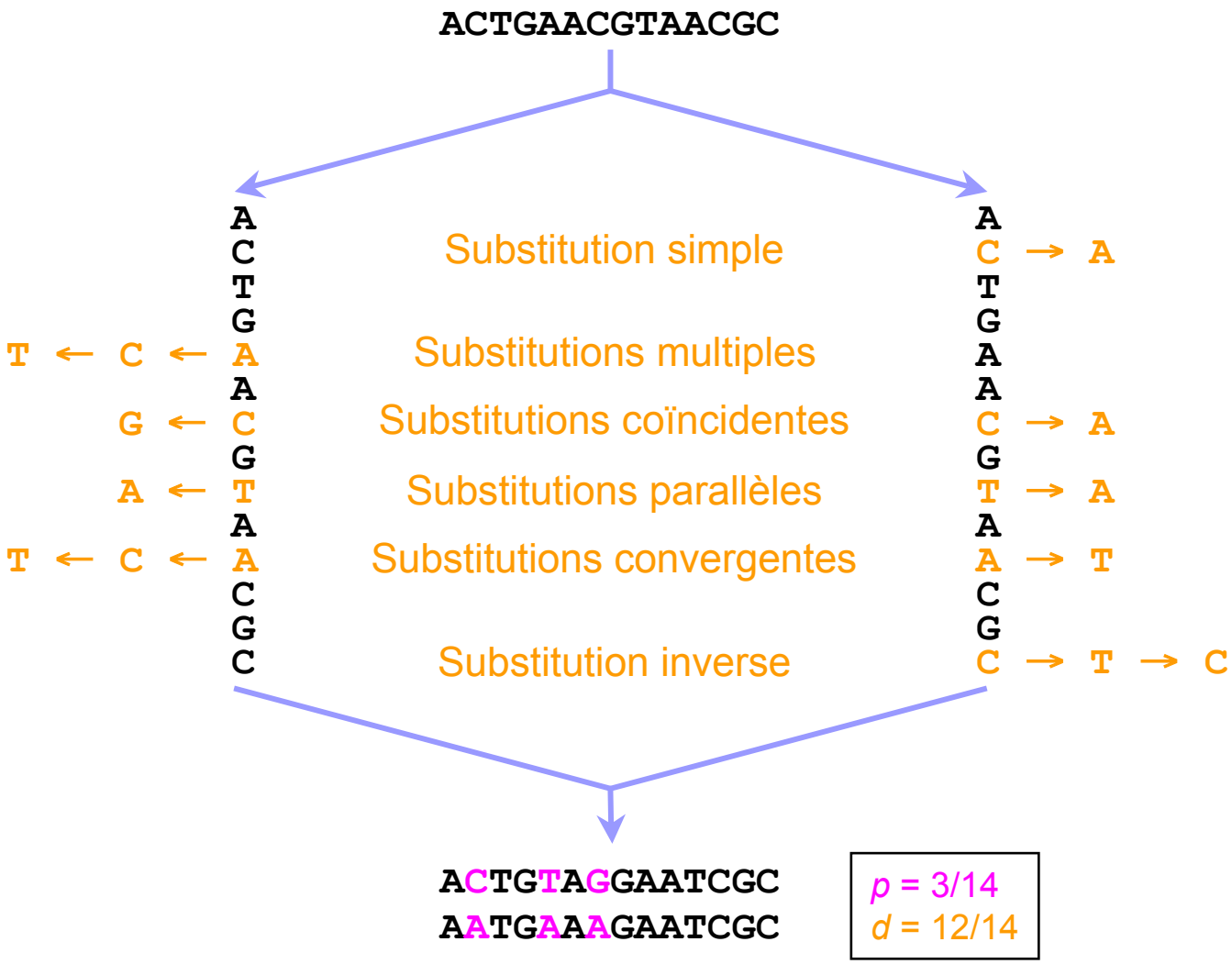
- Variation pour deux séquences de composition homogène :
  - Pour l'ADN : $0 \leq p \leq 0.75$.
  - Pour les protéines : $0 \leq p \leq 0.95$.

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.

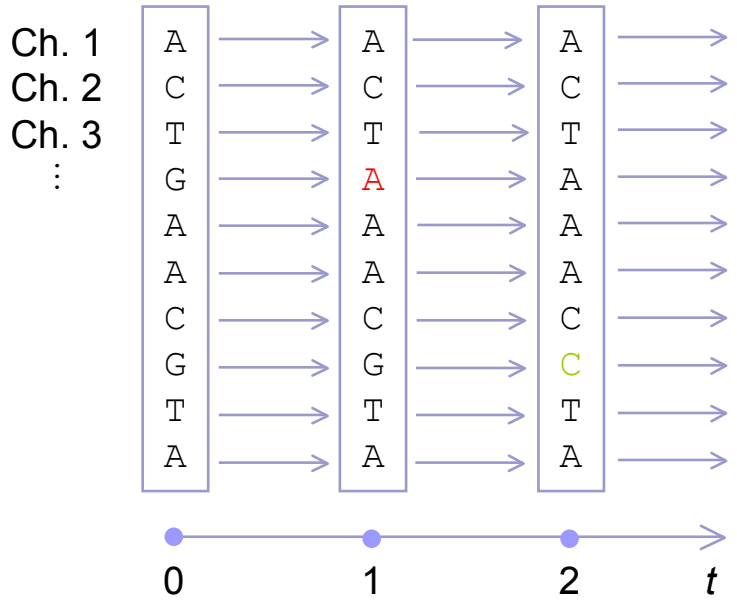


Types de substitutions



Modèles de Markov en phylogénie

- Utilisés pour les séquences nucléotidiques et protéiques.
- Les substitutions se font suivant un *processus de Markov*.
- Impliquent de déterminer des *probabilités de substitution* :
  - 16 valeurs en théorie pour les séquences d'ADN.
  - Moins en pratique :
    - Hypothèses simplificatrices.



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

Propriétés des modèles standards

- Temps continu.
- Hypothèses communes (simplificatrices) :
  - Stationnarité :
    - La fréquences des nucléotides/acides aminés dans les séquences est la même de la racine aux feuilles de l'arbre.
  - Réversibilité :
    - La quantité de changement d'un nucléotide/acide aminé $i \rightarrow j$ est égale à la quantité de changement $j \rightarrow i$.
  - Homogénéité par branche :
    - Un seul taux global de substitution tout au long de l'arbre.
  - Homogénéité par sites (ou uniformité) :
    - Tous les sites évoluent suivant le même processus.

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

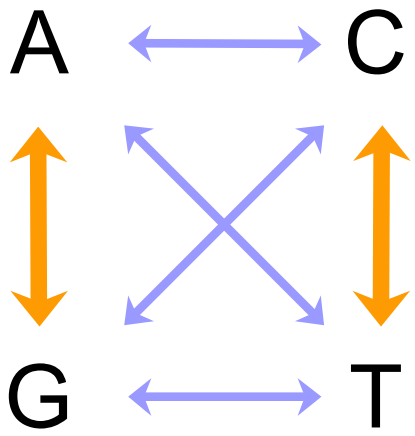
$$\mathbf{F} = \begin{pmatrix} f_{\mathrm{AA}} & f_{\mathrm{AC}} & f_{\mathrm{AT}} & f_{\mathrm{AG}} \\ f_{\mathrm{CA}} & f_{\mathrm{CC}} & f_{\mathrm{CT}} & f_{\mathrm{CG}} \\ f_{\mathrm{TA}} & f_{\mathrm{TC}} & f_{\mathrm{TT}} & f_{\mathrm{TG}} \\ f_{\mathrm{GA}} & f_{\mathrm{GC}} & f_{\mathrm{GT}}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

- Soient 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 :

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

Probabilités de substitution

- Soit $p_{ij}(t)$ la probabilité de substitution d'un nucléotide i vers un nucléotide j au cours du temps t :
  - *Probabilités de transition* du processus de Markov.
- L'ensemble de ces probabilités peuvent être regroupées dans une matrice $\mathbf{P}(t) = (p_{ij}(t))$ 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.

Taux de substitution

- Soit q_{ij} $(i \neq j)$ le *taux de substitution instantané* d'un nucléotide i vers un nucléotide j.
- Dans ce cas, le *taux de changement instantané* d'un nucléotide i est défini comme $\lambda_i = q_{ii} = \sum_{j \neq i} q_{ij}$.
- 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.

Calcul de $\mathbf{P}(t)$

- La dynamique des probabilités de substitutions pour un accroissement infinitésimal $\mathrm{d}t$ peut s'exprimer sous la forme :

$$\mathbf{P}(t + \mathrm{d}t) = \mathbf{P}(t)(\mathbf{I} + \mathbf{Q}\mathrm{d}t)$$

  avec $\mathbf{I}$ la matrice identité.
- La solution de l'équation précédente est :

$$\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$$

  Ce calcul nécessite la *diagonalisation* de la matrice $\mathbf{Q}$.

Stationnarité

- Au bout d'un temps infini, un processus de Markov atteint ce qu'on appelle un *état stationnaire* :
  - Les fréquences des différents états ne changent plus :

$$\lim_{t \to \infty} p_{ij}(t) = \pi_j$$

    avec π_j la *fréquence à l'équilibre* du nucléotide j.
- Les modèles standards considèrent que la stationnarité est atteinte dès la racine de l'arbre :
  - Utilisation des fréquences des bases dans le jeu de données pour estimer les valeurs des fréquences à l'équilibre.

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 $j \to i$.
- Pas de directionalité dans l'écoulement du temps :
  - Pas d'influence de la position de racine sur le calcul.

Échangeabilités

- En remplaçant les probabilités de substitution par les taux instantanés, l'équation précédente devient :

$$\pi_i \, q_{ij} = \pi_j \, q_{ji}$$

Soit :

$$\frac{q_{ij}}{\pi_j} = \frac{q_{ji}}{\pi_i} = s_{ij} = s_{ji}$$

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 :

$$\mathbf{Q} = \mathbf{S\Pi} = \begin{pmatrix} \cdot & s_{\mathrm{AC}} & s_{\mathrm{AT}} & s_{\mathrm{AG}} \\ s_{\mathrm{CA}} & \cdot & s_{\mathrm{CT}} & s_{\mathrm{CG}} \\ s_{\mathrm{TA}} & s_{\mathrm{TC}} & \cdot & s_{\mathrm{TG}} \\ s_{\mathrm{GA}} & s_{\mathrm{GC}} & s_{\mathrm{GT}} & \cdot \end{pmatrix} \times \begin{pmatrix} \pi_{\mathrm{A}} & 0 & 0 & 0 \\ 0 & \pi_{\mathrm{C}} & 0 & 0 \\ 0 & 0 & \pi_{\mathrm{T}} & 0 \\ 0 & 0 & 0 & \pi_{\mathrm{G}} \end{pmatrix}$$

avec $\mathbf{S} = (s_{ij})$ la matrice des échangeabilités et $\mathbf{\Pi} = \mathrm{diag}(\pi_i)$ la matrice diagonale des fréquences à l'équilibre.

Simplification de l'écriture

- Les échangeabilités étant symétriques, on pose :

$$s_{\mathrm{AC}} = s_{\mathrm{CA}} = \alpha \quad s_{\mathrm{AT}} = s_{\mathrm{TA}} = \beta$$
$$s_{\mathrm{AG}} = s_{\mathrm{GA}} = \gamma \quad s_{\mathrm{CT}} = s_{\mathrm{TC}} = \delta$$
$$s_{\mathrm{CG}} = s_{\mathrm{GC}} = \epsilon \quad s_{\mathrm{TG}} = s_{\mathrm{GT}} = \eta$$

- Et le produit matriciel précédent peut s'écrire :

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

Expression de Q

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

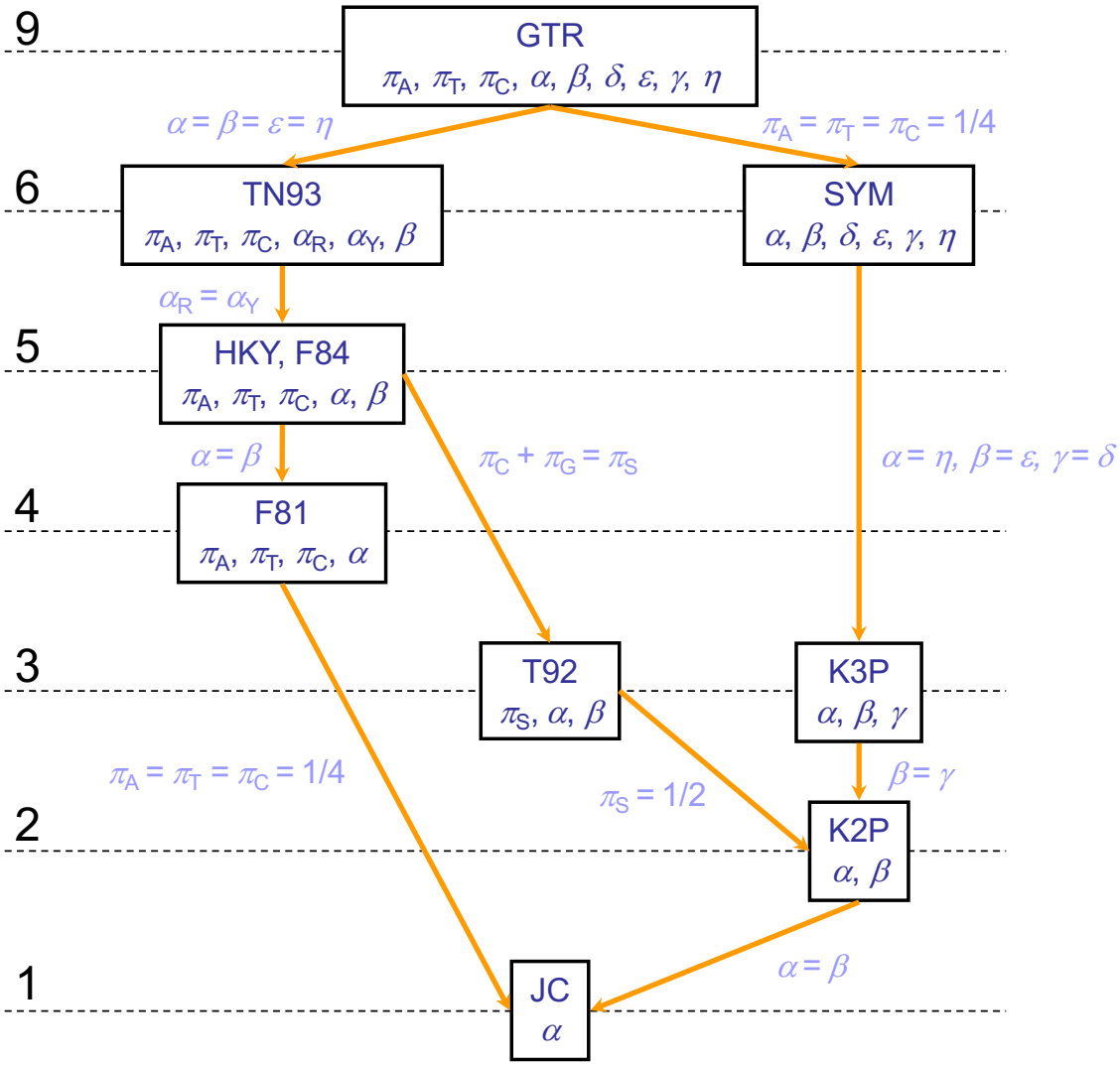
$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{C}}\alpha & \pi_{\mathrm{T}}\beta & \pi_{\mathrm{G}}\gamma \\ \pi_{\mathrm{A}}\alpha & -\lambda_{\mathrm{C}} & \pi_{\mathrm{T}}\delta & \pi_{\mathrm{G}}\epsilon \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\delta & -\lambda_{\mathrm{T}}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

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

Soit neuf paramètres à estimer :

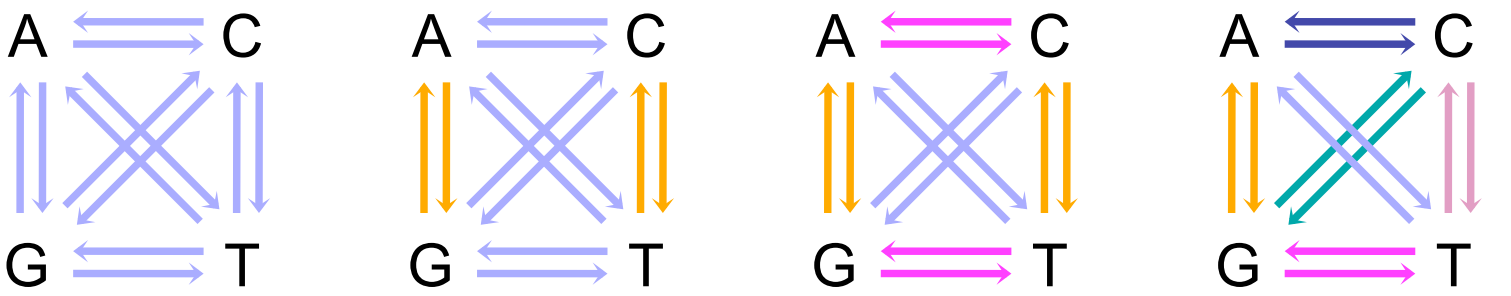
- Modèle GTR (*Generalised Time Reversible*) ou REV.

Imbrication des modèles

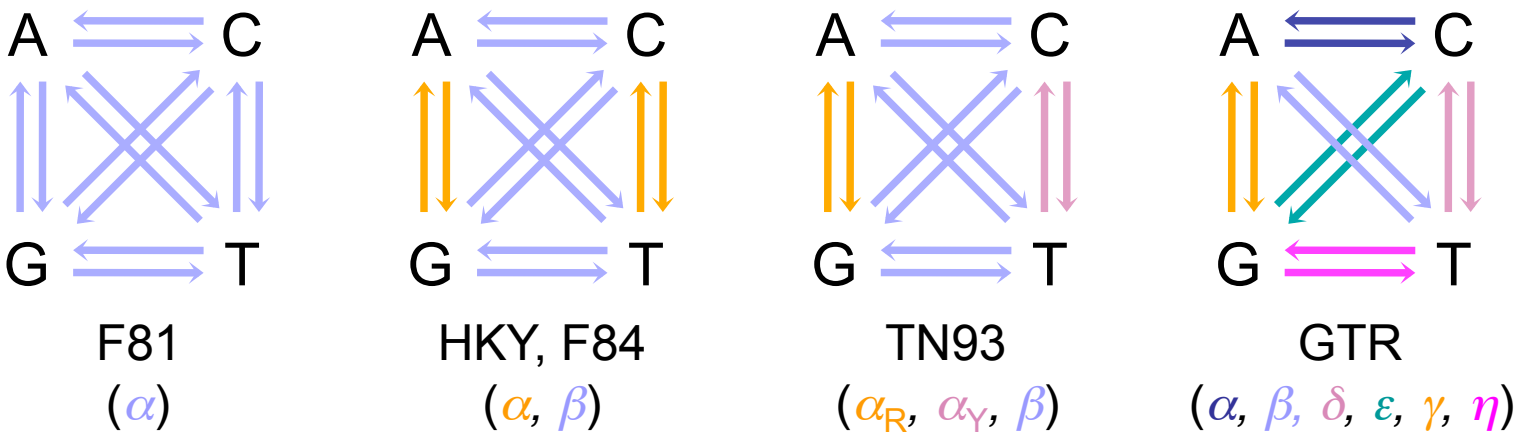


Paramètres des modèles

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



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



Calcul de la distance évolutive

- Soit λ, le *taux global de substitutions* entre deux séquences. Sous l'hypothèse de réversibilité, ce taux est égal à :

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

  avec $\lambda_i = \sum_{j \neq i} q_{ij}$ 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 :

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

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 :

$$d = \lambda t = t$$

Équivalence de ces deux expressions dans les notations utilisées par la suite.

Modèle de Jukes et Cantor

- Une seule échangeabilité (α), identique pour chacun des quatre nucléotides.
- Fréquences à l'équilibre $\pi_{\mathrm{A}} = \pi_{\mathrm{C}} = \pi_{\mathrm{T}} = \pi_{\mathrm{G}} = 1/4$.
- Matrice des taux instantanés :

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \alpha & \alpha & \alpha \\ \alpha & -\lambda & \alpha & \alpha \\ \alpha & \alpha & -\lambda & \alpha \\ \alpha & \alpha & \alpha & -\lambda \end{pmatrix}$$

- Taux global de substitutions :

$$\lambda = \sum_i \pi_i \lambda_i = 3\alpha = 1$$

Résolution

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

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4\alpha t} & (i = j) \\ \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\alpha t} & (i \neq j) \end{cases}$$

- Simplification en sachant que $3\alpha = 1$.
- Introduction de la divergence observée entre deux séquences $p = 3p_{ij}(t)\ (i \neq j)$.
- Formule de Jukes et Cantor pour le calcul de la distance évolutive :

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

Caractéristiques du modèle

- Variance de l'estimation :

$$\mathbb{V}(d) = \frac{9p(1-p)}{(3-4p)^2\ell}$$

- Ratio transitions/transversions :

$$\kappa = \frac{\alpha}{\alpha} = 1$$

  Ceci quelle que soit la composition des séquences.

- Limites d'utilisation :
  - Lorsque $p \to 3/4$, $d \to \infty$.
  - Le modèle n'est pas utilisable pour des séquences divergentes à plus de 75%.

Modèle de Kimura à deux paramètres

- L'échangeabilité pour les transitions (α) est différente de celle des transversions (β).
- Fréquences à l'équilibre $\pi_{\mathrm{A}} = \pi_{\mathrm{C}} = \pi_{\mathrm{T}} = \pi_{\mathrm{G}} = 1/4$.
- Matrice des taux instantanés :

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \beta & \beta & \alpha \\ \beta & -\lambda & \alpha & \beta \\ \beta & \alpha & -\lambda & \beta \\ \alpha & \beta & \beta & -\lambda \end{pmatrix}$$

- Taux global de substitutions :

$$\lambda = \sum_i \pi_i \lambda_i = \alpha + 2\beta = 1$$

Résolution

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

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}\mathrm{e}^{-4\beta t} + \frac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t} & (i = j) \\ \frac{1}{4} + \frac{1}{4}\mathrm{e}^{-4\beta t} - \frac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t} & (i \neq j, \text{ transition}) \\ \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\beta t} & (i \neq j, \text{ transversion}) \end{cases}$$

Calcul de la distance

- Simplification en sachant que $\alpha + 2\beta = 1$.
- Introduction de la fréquence des transitions et des transversions observées entre deux séquences :

$$r = p_{ij}(t) \quad (i \neq j, \text{ transition})$$
$$v = 2p_{ij}(t) \quad (i \neq j, \text{ transversion})$$

- Formule de Kimura pour le calcul du nombre de substitutions :

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

Caractéristiques du modèle

- Variance de l'estimation :

$$\mathbb{V}(d) = \frac{1}{\ell}\left[c_1^2 r + c_3^2 v - (c_1 r + c_3 v)^2\right]$$

avec $c_1 = 1/(1-2r-v)$, $c_2 = 1/(1-2v)$ et $c_3 = (c_1 + c_2)/2$.

- Ratio transitions/transversions :

$$\kappa = \frac{\alpha}{\beta} = \frac{2\ln(1-2r-v)}{\ln(1-2v)} - 1$$

- Limites d'utilisation :
  - Lorsque $v \to 1/2$, $d \to \infty$.
  - Le modèle n'est pas utilisable pour des séquences présentant plus de 50% de transversions.

Modèle de Felsenstein (1981)

- Une seule échangeabilité (α), identique pour chacun des quatre nucléotides.
- Fréquences à l'équilibre $\pi_{\mathrm{A}} \neq \pi_{\mathrm{C}} \neq \pi_{\mathrm{T}} \neq \pi_{\mathrm{G}}$.
- Matrice des taux instantanés :

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

- Taux global de substitutions :

$$\lambda = \sum_i \pi_i \lambda_i = 2\alpha(\pi_{\mathrm{R}}\pi_{\mathrm{Y}} + \pi_{\mathrm{A}}\pi_{\mathrm{G}} + \pi_{\mathrm{C}}\pi_{\mathrm{T}}) = 1$$

avec $\pi_{\mathrm{R}} = \pi_{\mathrm{A}} + \pi_{\mathrm{G}}$ et $\pi_{\mathrm{Y}} = \pi_{\mathrm{C}} + \pi_{\mathrm{T}}$.

Résolution

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

$$p_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j)\mathrm{e}^{-\alpha t} & (i = j) \\ \pi_j(1 - \mathrm{e}^{-\alpha t}) & (i \neq j) \end{cases}$$

- Simplification en sachant que $2\alpha(\pi_\mathrm{R}\pi_\mathrm{Y} + \pi_\mathrm{A}\pi_\mathrm{G} + \pi_\mathrm{C}\pi_\mathrm{T}) = 1$.
- Formule de Felsenstein (1981) pour le calcul du nombre de substitutions :

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

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

Caractéristiques du modèle

- Variance de l'estimation :

$$\mathbb{V}(d) = \frac{p(1-p)}{(1-p/a)^2 \ell}$$

- Ratio transitions/transversions :

$$\kappa = \frac{\alpha}{\alpha} = 1$$

  Ceci quelle que soit la composition des séquences.
- Limites d'utilisation :
  - Lorsque $p \to a$, $d \to \infty$.

Modèle de Tamura et Nei

- Deux échangeabilités pour les transitions entre purines (α_{R}) et pyrimidines (α_{Y}) et une échangeabilité pour les transversions (β).
- Fréquences à l'équilibre $\pi_{\mathrm{A}} \neq \pi_{\mathrm{C}} \neq \pi_{\mathrm{T}} \neq \pi_{\mathrm{G}}$.
- Matrice des taux instantanés :

$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{T}}\beta & \pi_{\mathrm{G}}\alpha_{\mathrm{R}} \\ \pi_{\mathrm{A}}\beta & -\lambda_{\mathrm{C}} & \pi_{\mathrm{T}}\alpha_{\mathrm{Y}} & \pi_{\mathrm{G}}\beta \\ \pi_{\mathrm{A}}\beta & \pi_{\mathrm{C}}\alpha_{\mathrm{Y}} & -\lambda_{\mathrm{T}} & \pi_{\mathrm{G}}\beta \\ \pi_{\mathrm{A}}\alpha_{\mathrm{R}} & \pi_{\mathrm{C}}\beta & \pi_{\mathrm{T}}\beta & -\lambda_{\mathrm{G}} \end{pmatrix}$$

- Taux global de substitutions :

$$\lambda = \sum_i \pi_i \lambda_i = 2(\pi_{\mathrm{A}}\pi_{\mathrm{G}}\alpha_{\mathrm{R}} + \pi_{\mathrm{T}}\pi_{\mathrm{C}}\alpha_{\mathrm{Y}} + \pi_{\mathrm{R}}\pi_{\mathrm{Y}}\beta) = 1$$

Résolution

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

$$p_{ij}(t) = \begin{cases} \pi_j(1 - \mathrm{e}^{-\beta t}) & (1) \\ \pi_j + \dfrac{\pi_j \pi_{\mathrm{R}}}{\pi_{\mathrm{Y}}} \mathrm{e}^{-\beta t} - \dfrac{\pi_j}{\pi_{\mathrm{Y}}} \mathrm{e}^{-(\pi_{\mathrm{Y}}\alpha_{\mathrm{Y}} + \pi_{\mathrm{R}}\beta)t} & (2) \\ \pi_j + \dfrac{\pi_j \pi_{\mathrm{Y}}}{\pi_{\mathrm{R}}} \mathrm{e}^{-\beta t} - \dfrac{\pi_j}{\pi_{\mathrm{R}}} \mathrm{e}^{-(\pi_{\mathrm{R}}\alpha_{\mathrm{R}} + \pi_{\mathrm{Y}}\beta)t} & (3) \\ \pi_j + \dfrac{\pi_j \pi_{\mathrm{R}}}{\pi_{\mathrm{Y}}} \mathrm{e}^{-\beta t} + \dfrac{\pi_k}{\pi_{\mathrm{Y}}} \mathrm{e}^{-(\pi_{\mathrm{Y}}\alpha_{\mathrm{Y}} + \pi_{\mathrm{R}}\beta)t} & (4) \\ \pi_j + \dfrac{\pi_j \pi_{\mathrm{Y}}}{\pi_{\mathrm{R}}} \mathrm{e}^{-\beta t} + \dfrac{\pi_k}{\pi_{\mathrm{R}}} \mathrm{e}^{-(\pi_{\mathrm{R}}\alpha_{\mathrm{R}} + \pi_{\mathrm{Y}}\beta)t} & (5) \end{cases}$$

(1) Transversions
(2) $i, j \in \{\mathrm{C}, \mathrm{T}\}$ et $i \neq j$
(3) $i, j \in \{\mathrm{A}, \mathrm{G}\}$ et $i \neq j$
(4) $i, j, k \in \{\mathrm{C}, \mathrm{T}\}$, $i = j$ et $k \neq j$
(5) $i, j, k \in \{\mathrm{A}, \mathrm{G}\}$, $i = j$ et $k \neq j$

Calcul de la distance

- Simplification en sachant que $2(\pi_A\pi_G\alpha_R + \pi_T\pi_C\alpha_Y + \pi_R\pi_Y\beta) = 1$.
- Formule de Tamura et Nei pour le calcul du nombre de substitutions :

$$d = \frac{2\pi_T\pi_C}{\pi_Y}(a_1 - \pi_R b) + \frac{2\pi_A\pi_G}{\pi_R}(a_2 - \pi_Y b) + 2\pi_Y\pi_R b$$

$$\text{avec } \begin{cases} a_1 = -\ln\left(1 - \dfrac{\pi_Y}{2\pi_T\pi_C}r_Y - \dfrac{1}{2\pi_Y}v\right) \\ a_2 = -\ln\left(1 - \dfrac{\pi_R}{2\pi_A\pi_G}r_R - \dfrac{1}{2\pi_R}v\right) \\ b = -\ln\left(1 - \dfrac{1}{2\pi_R\pi_Y}v\right) \end{cases}$$

Caractéristiques du modèle

- Variance de l'estimation :
  - Trop complexe pour tenir sur une seule diapositive (*cf.* Tamura et Nei, 1993) !
- Ratio transitions/transversions :

$$\kappa_{\mathrm{Y}} = \frac{\alpha_{\mathrm{Y}}}{\beta} = \frac{a_1 - \pi_{\mathrm{R}} b}{\pi_{\mathrm{Y}} b} \quad \text{et} \quad \kappa_{\mathrm{R}} = \frac{\alpha_{\mathrm{R}}}{\beta} = \frac{a_2 - \pi_{\mathrm{Y}} b}{\pi_{\mathrm{R}} b}$$

  avec $\kappa = \kappa_{\mathrm{Y}} + \kappa_{\mathrm{R}}$.
- Limites d'utilisation :
  - Logarithmes indéfinis pour les valeurs de a_1, a_2 ou b si séquences trop divergentes (valeurs négatives).
  - Le modèle n'est pas utilisable pour des séquences présentant plus de 50% de transversions.

Modèle GTR

- Pas de solution analytique au calcul de $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$.
- Expression matricielle du calcul de la distance :
  - Soit $\mathbf{\Pi} = \mathrm{diag}(\pi_i)$ la matrice diagonale contenant les valeurs des fréquences des bases à l'équilibre, dans ce cas on a :

$$d = \lambda t = -\operatorname{trace}(\mathbf{\Pi Q} t)$$

    Du fait que $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$, l'expression ci-dessus est équivalente à :

$$d = -\operatorname{trace}[\mathbf{\Pi} \ln \mathbf{P}(t)]$$

  - Estimation de $\mathbf{\Pi}$ à partir des fréquences des bases dans le jeu de données.
  - Quel estimateur pour $\mathbf{P}(t)$?

Introduction de $\mathbf{F}(t)$

- Soit $f_{ij}(t)$ la probabilité d'avoir au temps t une substitution $i \rightarrow j$ dans l'alignement :

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

Soit, sous forme matricielle :

$$\mathbf{F}(t) = \mathbf{\Pi P}(t)$$

- Dans ce cas, la formule permettant le calcul de la distance peut s'écrire comme :

$$d = -\operatorname{trace}\left\{\mathbf{\Pi} \ln\left[\mathbf{\Pi}^{-1}\mathbf{F}(t)\right]\right\}$$

Du fait que $\mathbf{P}(t) = \mathbf{\Pi}^{-1}\mathbf{F}(t)$.

Estimation de $\mathbf{F}(t)$

- Utilisation de $\mathbf{F}$, construite à partir des fréquences des substitutions observées dans le jeu de données (*cf.* Diapo. 40).
- $\mathbf{\Pi}^{-1}\mathbf{F}$ doit être diagonalisable pour permettre le calcul de son logarithme.
- Les valeurs propres de $\mathbf{\Pi}^{-1}\mathbf{F}$ doivent être des réels positifs non nuls :
  - Symétrisation de $\mathbf{F}$ en $\mathbf{F}^*$ par l'opération :

$$\mathbf{F}^* = \frac{1}{2}\left(\mathbf{F} + \mathbf{F}^{\mathrm{T}}\right)$$

$\mathbf{\Pi}^{-1}\mathbf{F}^*$ correspond alors au produit d'une matrice diagonale par une matrice symétrique définie positive.

Utilité des modèles complexes

- Modélisent mieux l'évolution des séquences :
  - Plus proches de la réalité biologique.
  - Sous-estimation des distances évolutives par les modèles simples.
- Nécessitent de disposer de données en quantités plus importantes :
  - Nombre croissant de paramètres à estimer.
  - Augmentation de la variance avec le nombre de paramètres.
- Séquences trop divergentes :
  - Impossible de calculer d.
- Tests de sélection de modèles (maximum de vraisemblance) :
  - Compromis entre l'augmentation du nombre de paramètres et le gain de vraisemblance.

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 GTR pour les protéines ?

- Matrice 20×20 des taux instantanés :

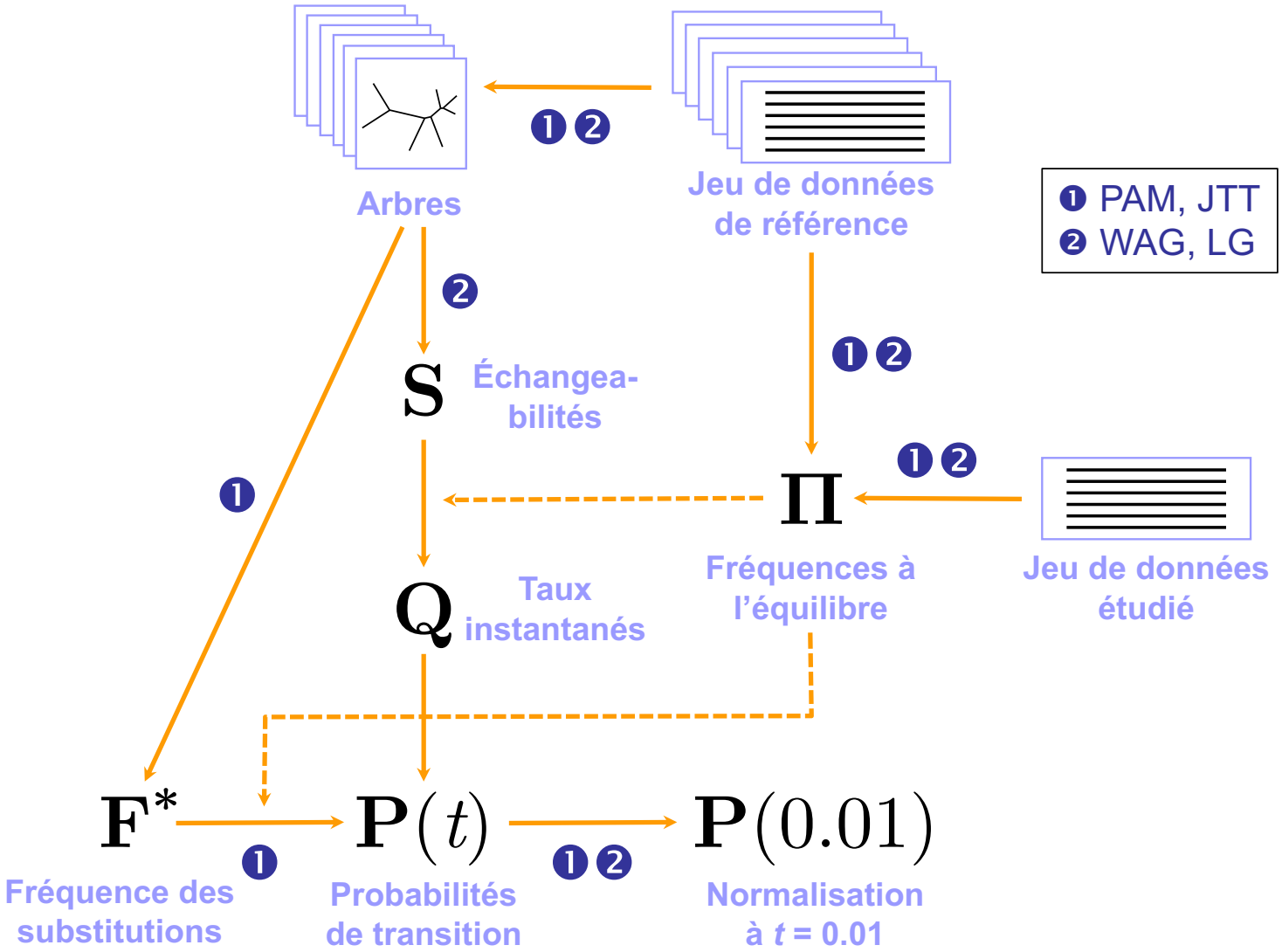
$$\mathbf{Q} = \mathbf{S\Pi} = \begin{pmatrix} -\lambda_{\mathrm{A}} & \pi_{\mathrm{R}} s_{\mathrm{AR}} & \cdots & \pi_{\mathrm{V}} s_{\mathrm{AV}} \\ \pi_{\mathrm{A}} s_{\mathrm{AR}} & -\lambda_{\mathrm{R}} & \cdots & \pi_{\mathrm{V}} s_{\mathrm{RV}} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{\mathrm{A}}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

  Soit 190 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 réalisable 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 référence comprenant un grand nombre de séquences.

Modèles empiriques courants

- Approche par comptage des substitutions :
  - PAM (*Point Accepted Mutation*, Dayhoff *et al.*, 1978) :
    - 1300 séquences provenant de 71 familles pour un total de 1572 substitutions.
    - Existence de plusieurs variantes (DCMut, Gonnet).
  - JTT (Jones, Taylor et Thornton, 1992) :
    - 16300 séquences totalisant 59190 substitutions.
- Approche au maximum de vraisemblance :
  - WAG (Whelan et Goldman, 2001) :
    - 3905 séquences provenant de 182 familles.
  - LG (Le et Gascuel, 2008) :
    - 49637 séquences provenant de 3912 familles.
  - Modèles « spécialisés » (mtMAM, cpREV, HIVb, FLU, etc.)

Calcul de la matrice $\mathbf{P}(t)$



Calcul d'une distance évolutive

- La divergence observée entre deux séquences est égale à :

$$p = \sum_{i \neq j} f_{ij}(t) = \sum_{i \neq j} \pi_i p_{ij}(t) = 1 - \sum_i \pi_i p_{ii}(t)$$

- Pour $p = 0.01$, on fait l'hypothèse de l'absence de substitutions multiples, ce qui implique :

$$t = p = 1 - \sum_i \pi_i p_{ii}(0.01) = 0.01$$

- Maintenant, si $p \neq 0.01$, comment calculer t ?

Propriété de Chapman-Kolmogorov

- Pour une chaîne de Markov en temps continu, la propriété de Chapman-Kolmogorov fait que :

$$\mathbf{P}(r \times t) = \mathbf{P}(t)^r \quad (r > 0)$$

- Toute distance peut donc s'exprimer sous la forme $t = r \times 0.01$, r étant la puissance à laquelle il faut élever $\mathbf{P}(0.01)$ de façon à ce que :

$$p = 1 - \sum_i \pi_i p_{ii}(r \times 0.01) = p_{\text{obs}}$$

avec p_{obs} la divergence observée entre les deux séquences.

- Calcul itératif jusqu'à convergence vers la bonne valeur.

Exemple numérique

- Calcul de t pour $p_{\text{obs}} = 0.17$ avec le modèle PAM :
  - $\mathbf{P}(0.01) = \mathbf{P}(0.01)^1 \Rightarrow p = 0.01 < p_{\text{obs}}$
  - $\mathbf{P}(0.17) = \mathbf{P}(0.01)^{17} \Rightarrow p \simeq 0.15201 < p_{\text{obs}}$
  - $\mathbf{P}(0.18) = \mathbf{P}(0.01)^{18} \Rightarrow p \simeq 0.15993 < p_{\text{obs}}$
  - $\mathbf{P}(0.19) = \mathbf{P}(0.01)^{19} \Rightarrow p \simeq 0.16774 < p_{\text{obs}}$
  - $\mathbf{P}(0.20) = \mathbf{P}(0.01)^{20} \Rightarrow p \simeq 0.17556 > p_{\text{obs}}$
  - $\mathbf{P}(0.195) = \mathbf{P}(0.01)^{19.5} \Rightarrow p \simeq 0.17162 > p_{\text{obs}}$
  - $\mathbf{P}(0.1925) = \mathbf{P}(0.01)^{19.25} \Rightarrow p \simeq 0.16968 < p_{\text{obs}}$
  - $\mathbf{P}(0.19375) = \mathbf{P}(0.01)^{19.375} \Rightarrow p \simeq 0.17065 > p_{\text{obs}}$
  - $\ldots$
  - $\mathbf{P}(0.19291) = \mathbf{P}(0.01)^{19.291} \Rightarrow p \simeq 0.17 \Rightarrow t \simeq 0.19291$
- Le même calcul avec un autre modèle donnerait un résultat différent.

Approximation des distances

- Calcul rapide d'une distance évolutive avec les ordinateurs actuels, mais pas au moment de la conception des modèles.
- Approximation de Kimura (1983) pour PAM :

$$\hat{t} = -\ln(1 - p - 0.2p^2)$$

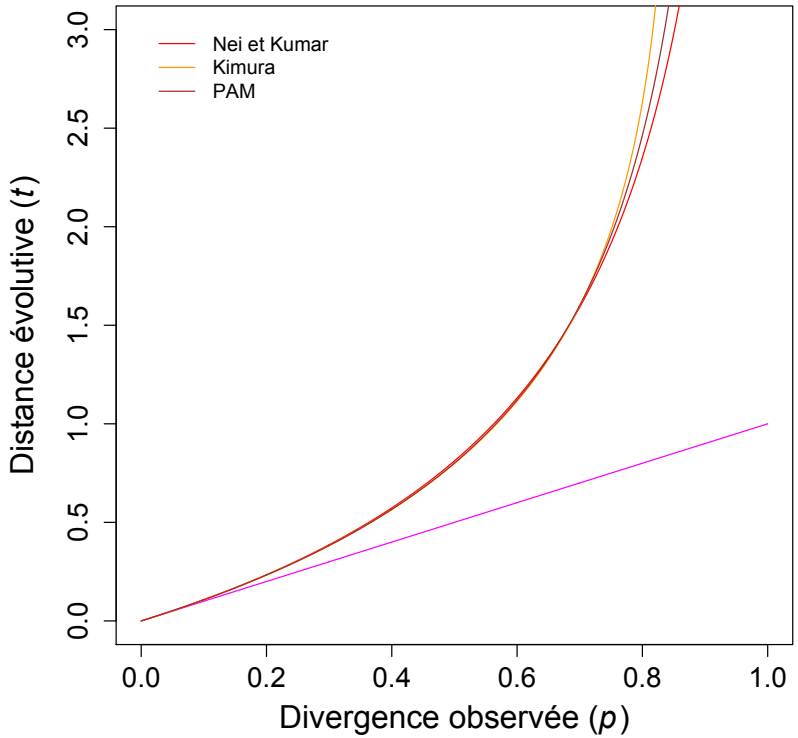
- Approximation de Nei et Kumar (2000) pour PAM et JTT :

$$\hat{t} = a[(1 - p)^{-1/a} - 1]$$

avec $a = 2.25$ (PAM) et $a = 2.4$ (JTT).

Limites d'utilisation

- Méthodes simples et rapides :
  - Encore implémentées dans des programmes actuels.
- Limitations :
  - Pas de prise en compte des fréquences à l'équilibre des séquences étudiées.
  - La précision de l'estimation diminue avec la divergence.



Matrices de scores

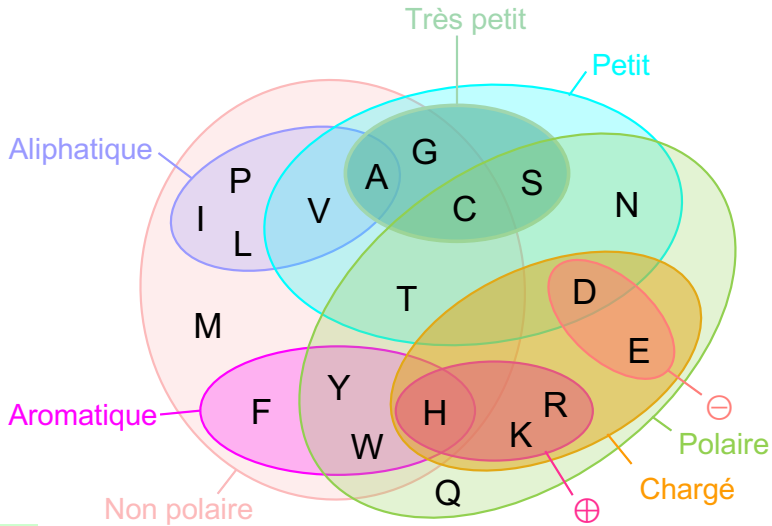
- Utilisées par les programmes d'alignement et de recherche de similarités (*e.g.*, BLAST).
- Fondées sur les matrices de probabilités de transition :
  - Calcul d'un ensemble de matrices obtenues par exponentiation de $\mathbf{P}(0.01)$ pour différentes valeurs de t (*e.g.*, 0.5, 1.5, 2.5).
  - Pour chacune des matrices $\mathbf{P}(t)$ précédentes, construction d'une matrice de score $\mathbf{M}(t) = (\mu_{ij}(t))$, telle que :

$$\mu_{ij}(t) = 10 \log \left(\frac{p_{ij}(t)}{\pi_j} \right)$$

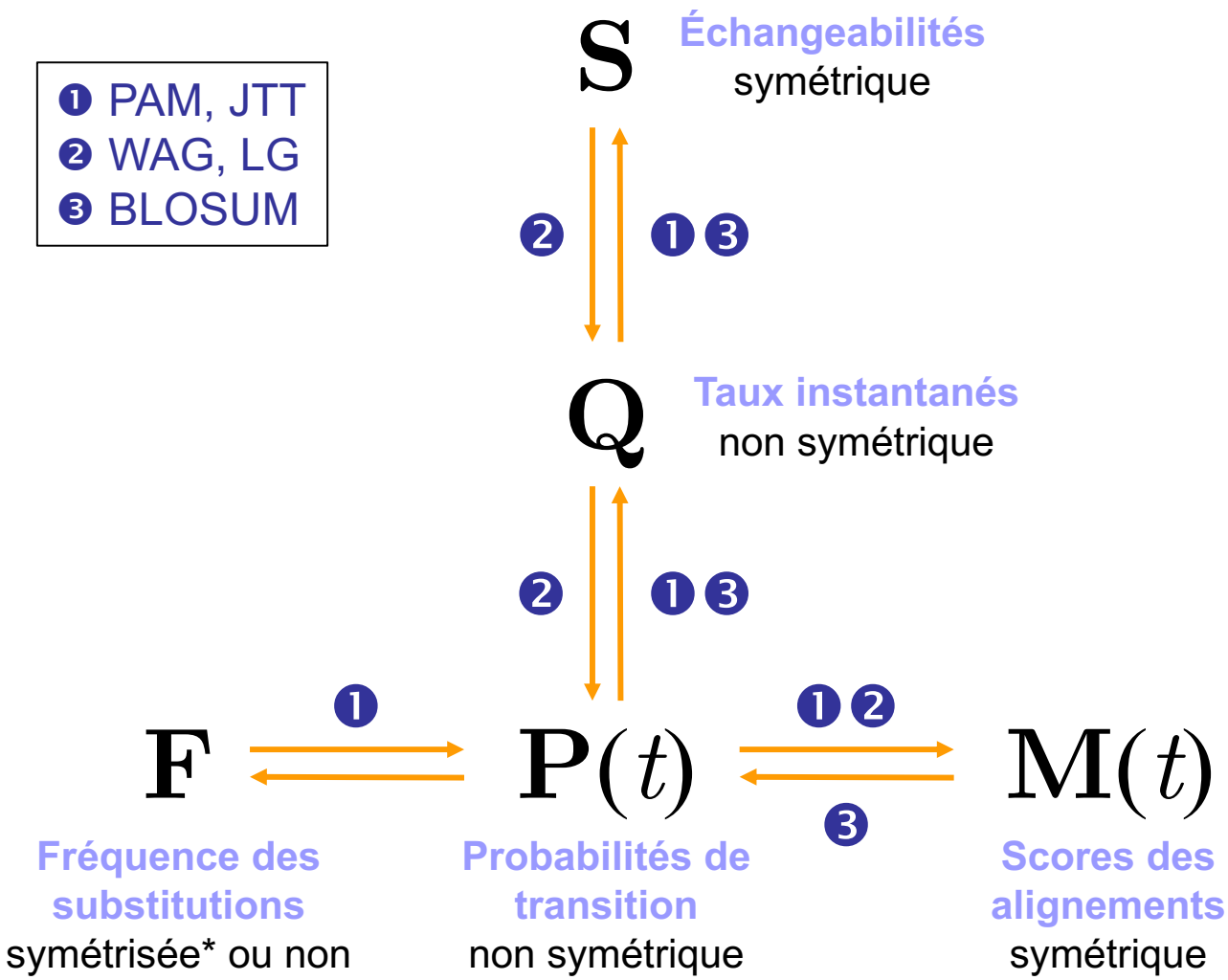
    avec arrondi à l'entier le plus proche.
  - Construction de PAM250 à partir de la matrice de probabilités de transition $\mathbf{P}(2.5) = \mathbf{P}(0.01)^{250}$ de PAM.

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	7											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-3	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4



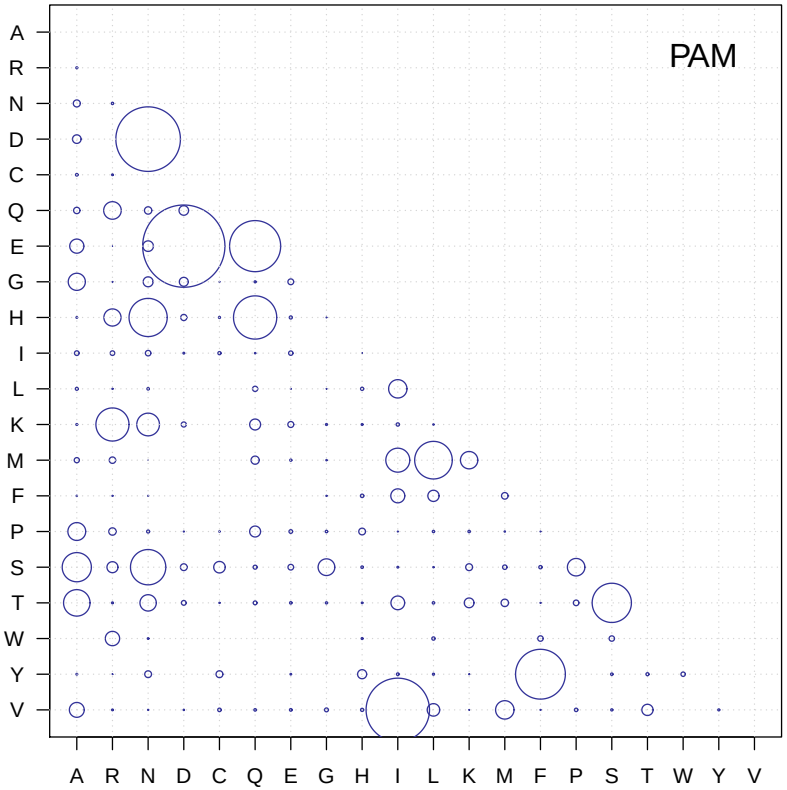
Interconversion des matrices

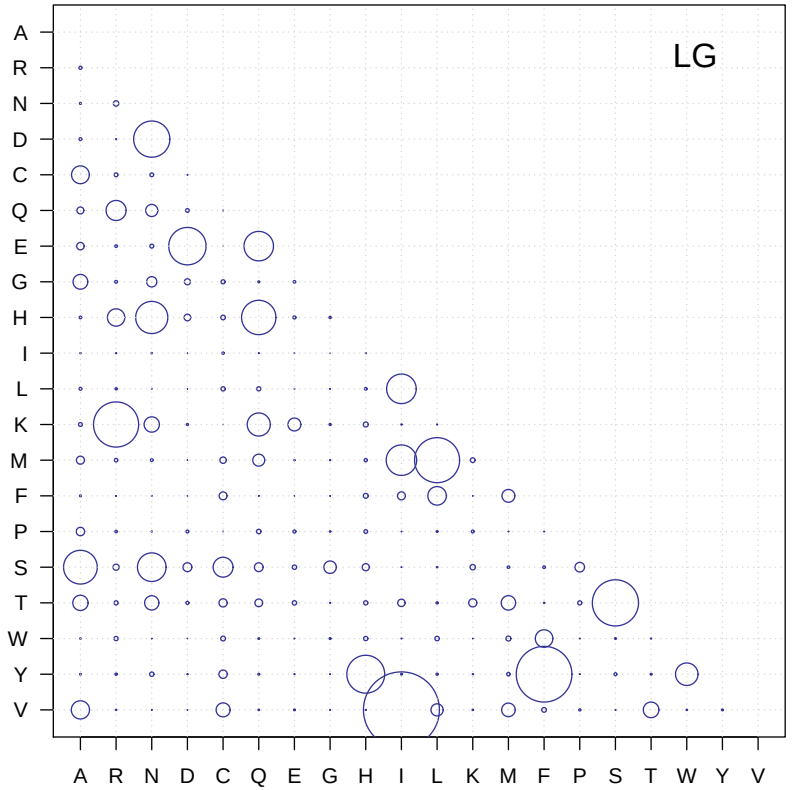


Comparaison des échangeabilités

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

Problème lié à la taille de l'échantillon utilisé



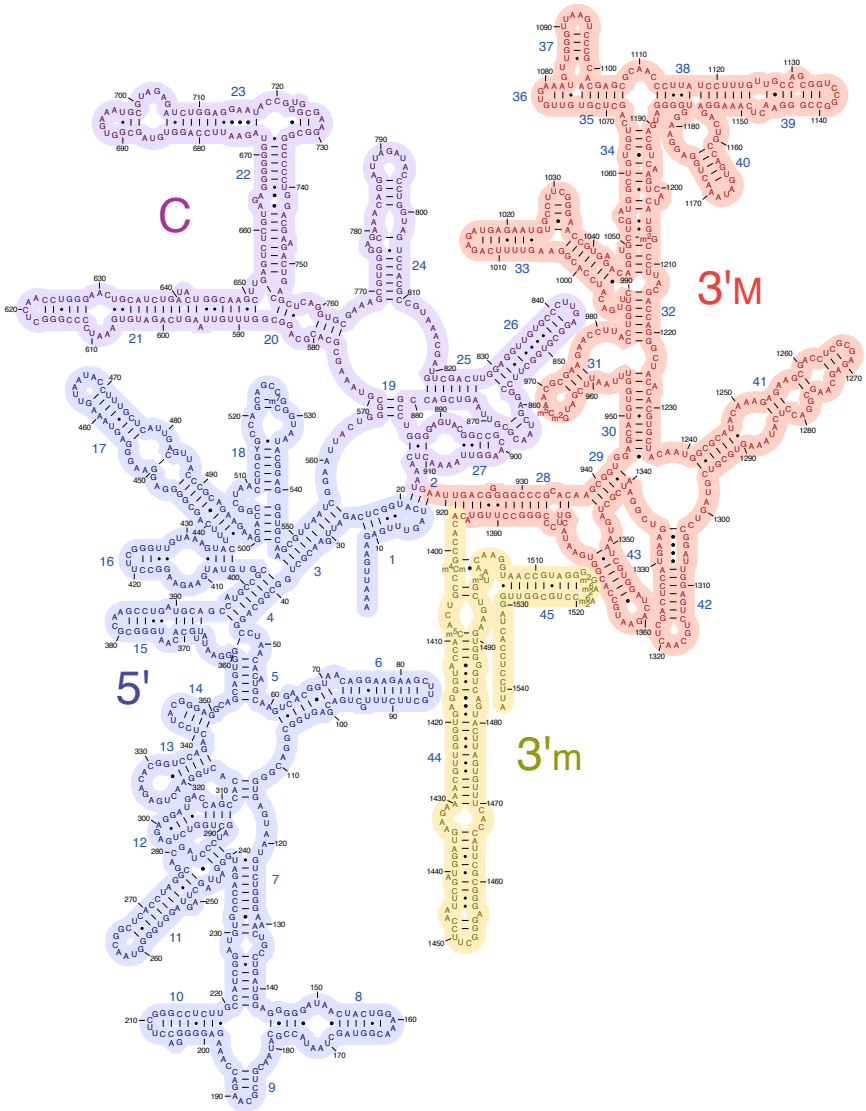


Modèles hétérogènes

- Les hypothèses d'homogénéité sont la plupart du temps non vérifiées :
  - Positions I, II et III des codons.
  - Contraintes structurales (ARNr, protéines).
  - Accélération évolutive dans certaines lignées.
- Utilisation de modèles *hétérogènes* :
  - Hétérogénéité par *sites* :
    - Correction par la loi Gamma.
    - Modèles de mélange.
    - Modèles de partition (concaténations).
  - Hétérogénéité par *branches*.
  - Hétérogénéité par sites et par branches.

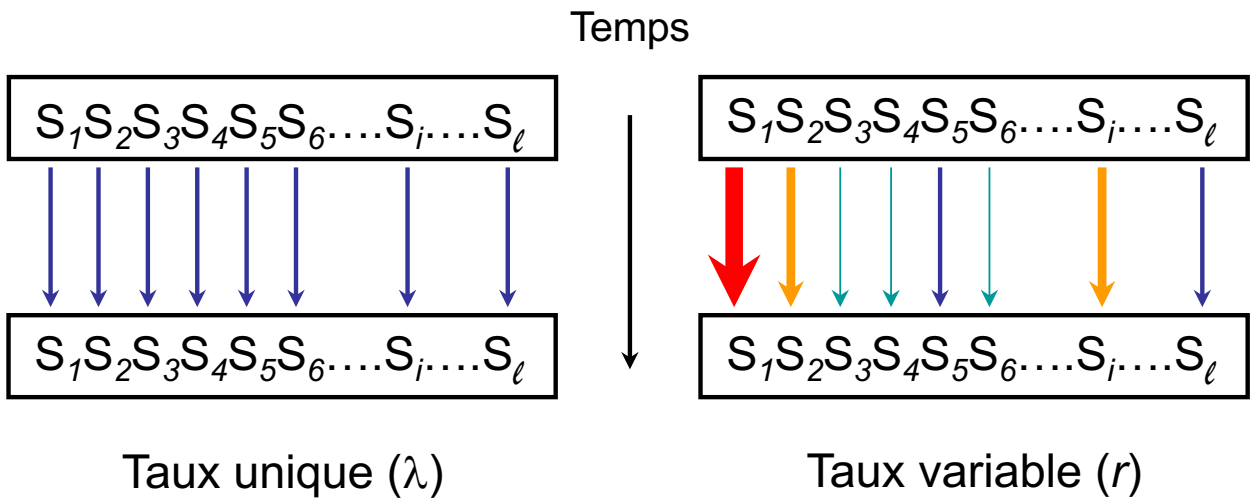
Exemple de l'ARNr 16S

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



Correction par la loi Gamma

- Hypothèse des modèles homogènes :
  - Tous les sites possèdent le même taux instantané de substitution dont la valeur normalisée est fixée à $\lambda = 1$.
- Proposition par Yang (1994) d'utiliser un taux variable r :
  - Tirage de la valeur de r dans une distribution Gamma.



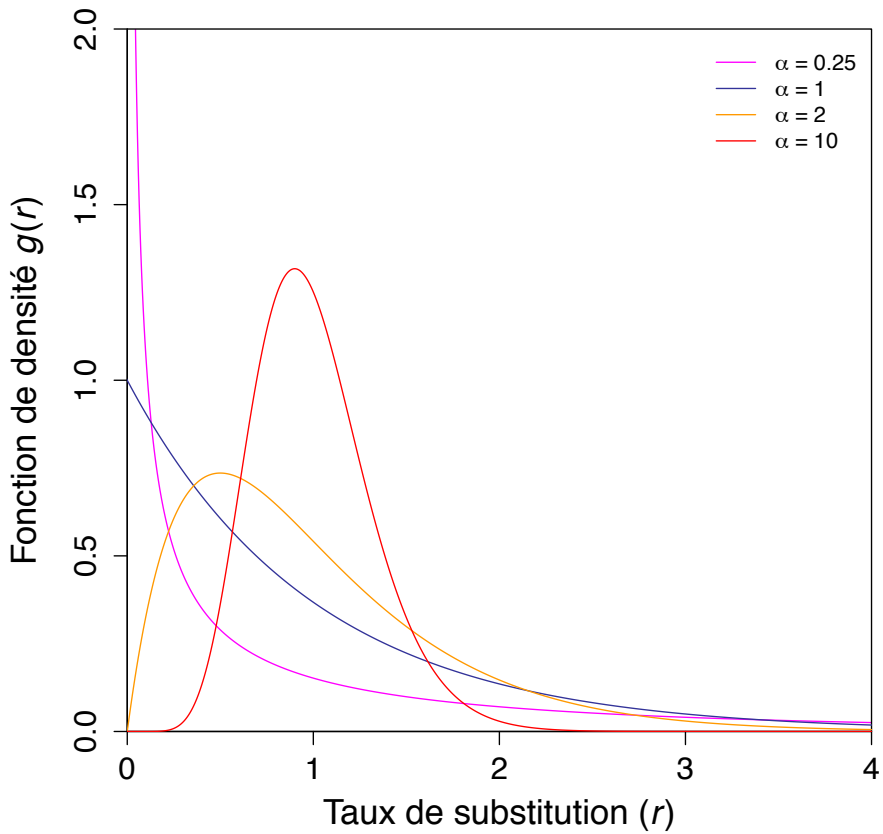
Distribution Gamma

- 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

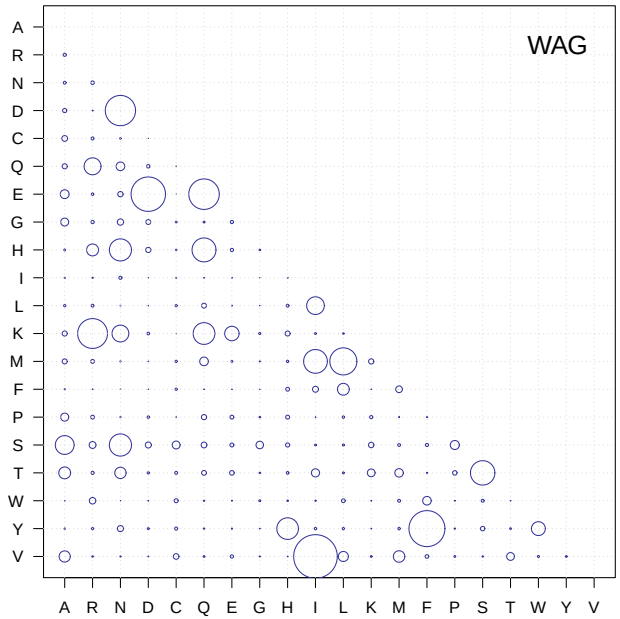
- Nombre de classes K fixé par l'utilisateur ($2 \leq K \leq 8$).
- Bornes z_k ($k = \{1, 2, \ldots, K-1\}$) correspondants aux quantiles à k/K de la distribution Gamma correspondante :
  - Le taux d'un site tiré au hasard a une probabilité $1/K$ d'appartenir à chacune d'entre elles.
- Ajout éventuel d'une classe supplémentaire pour prendre en compte les sites *invariants* :
  - Cas particulier où $r = 0$.

Décorations d'un modèle

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

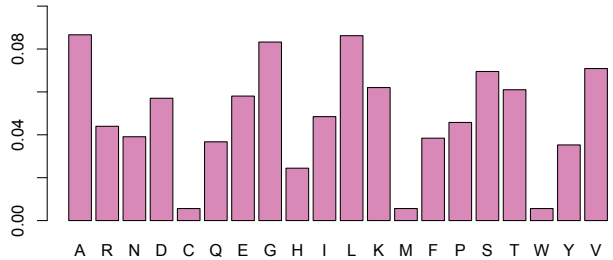
Approche classique

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



$\mathbf{S}$

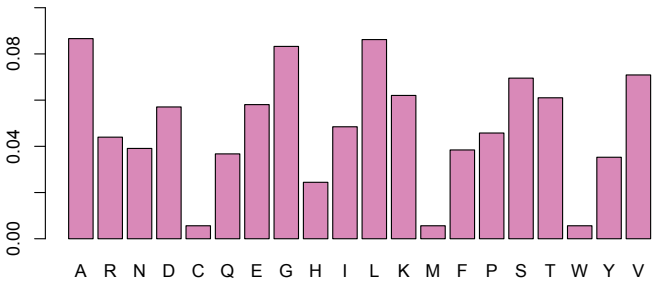
$\times$



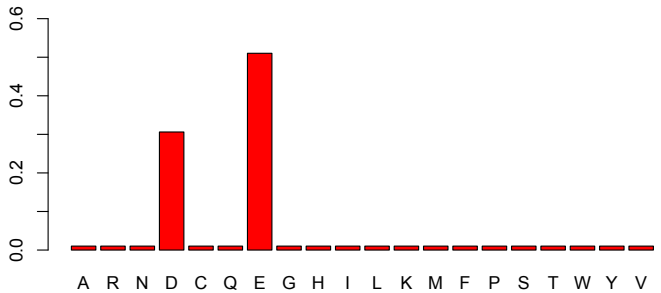
$\mathbf{\Pi}$

$= \mathbf{Q}$

Limites de l'approche classique



Fréquences à l'équilibre globales
(peu adaptées)

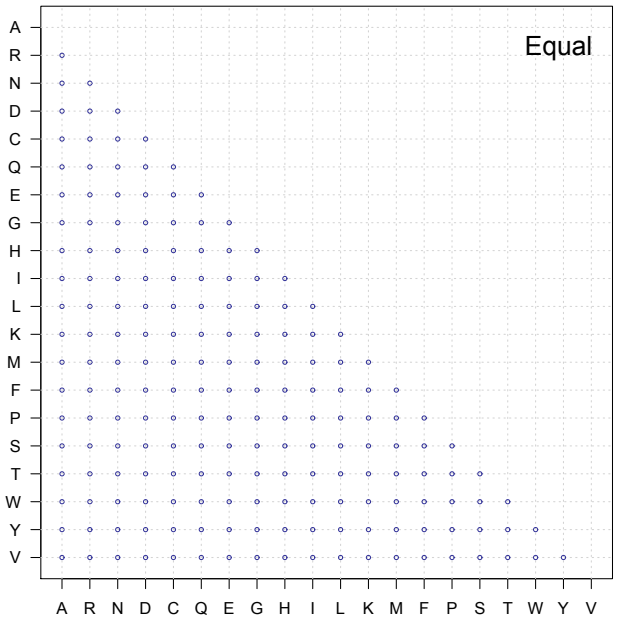


Fréquences à l'équilibre site spécifiques
(plus réalistes)

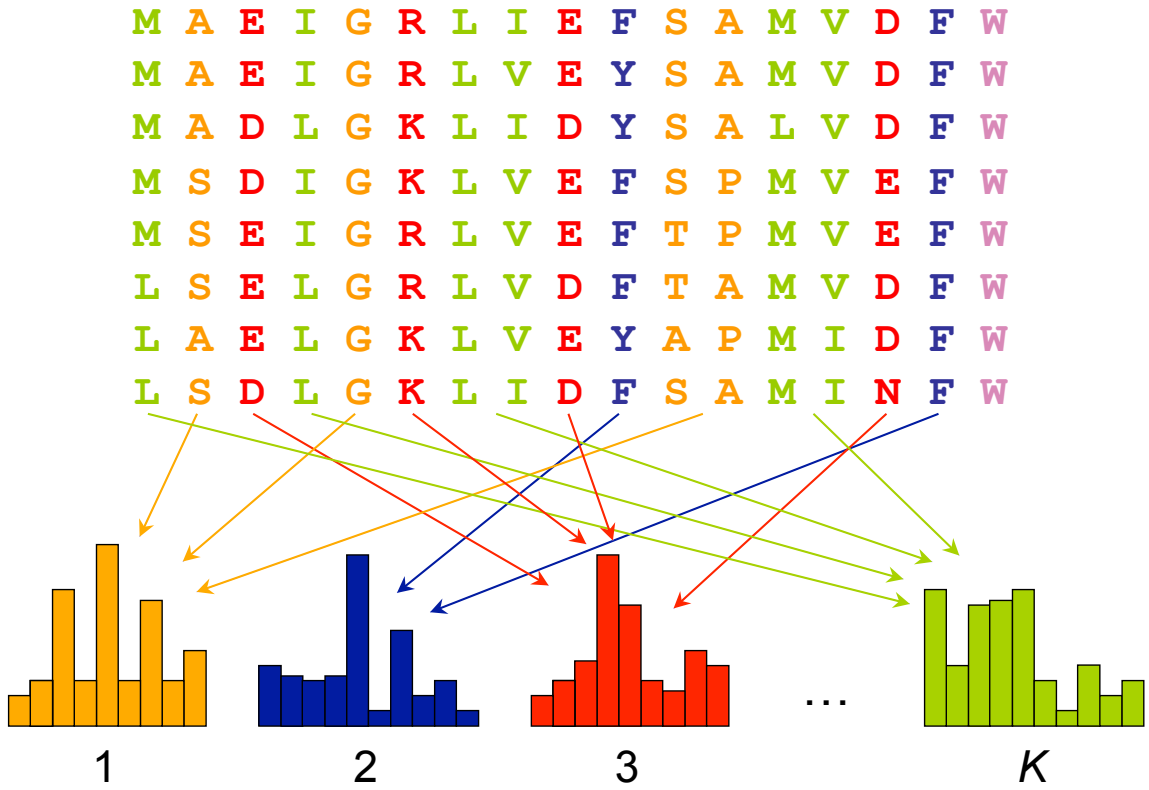
Modèles de mélange

- L'utilisation d'un jeu de valeurs π_i unique n'est pas 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

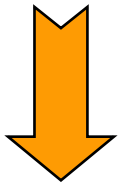


K catégories de valeurs de π_i
(K = 20, 30, 40, 50, 60)

Plan

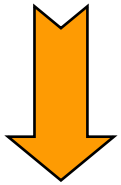
Principe général

Alignement de séquences



Mesures de distances évolutives

Matrice de distances évolutives entre paires de séquences

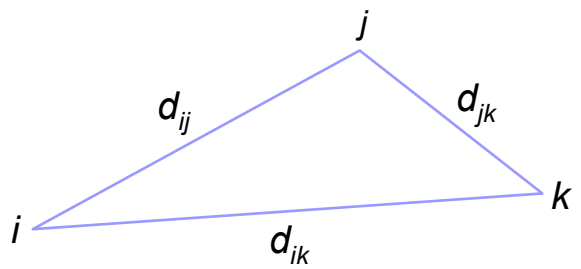


Calcul de l'arbre à partir de la matrice

Arbre

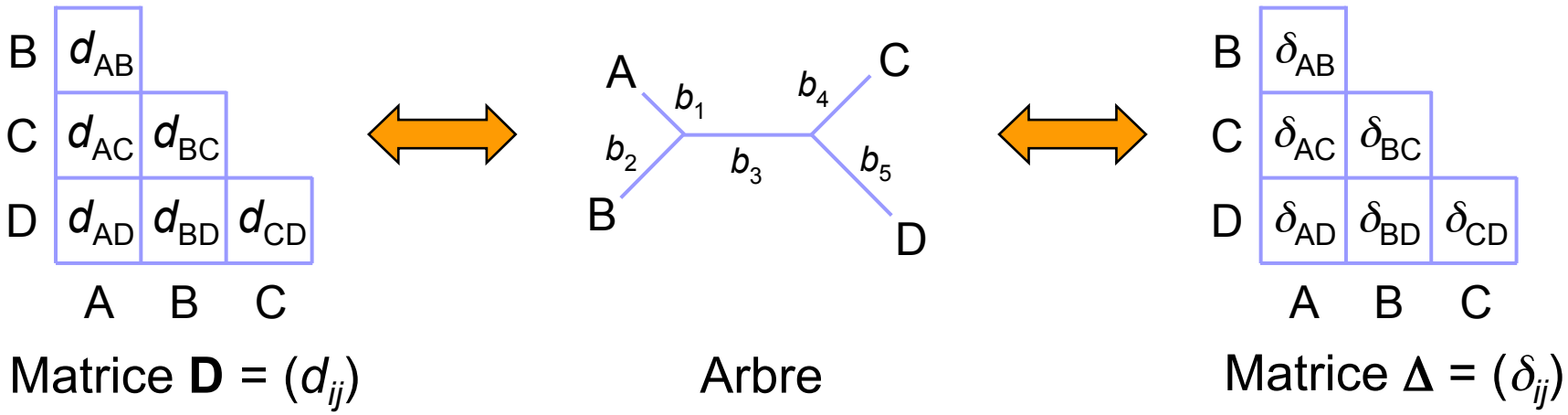
Notion de distance

- En mathématiques, une distance (ou *métrique*) sur un ensemble E est une fonction $d : E \times E \mapsto \mathbb{R}^+$.
- Cette fonction doit satisfaire à trois conditions, ceci $\forall i, j, k \in E$:
  - *Symétrie* – la distance entre deux points est la même, quelle que soit la direction considérée ($d_{ij} = d_{ji}$).
  - *Séparation* – si la distance entre deux points est égale à zéro, alors ces deux points sont confondus ($d_{ij} = 0 \Leftrightarrow i = j$).
  - *Inégalité triangulaire* – le chemin direct entre deux points est le plus court ($d_{ik} \leq d_{ij} + d_{jk}$) :



Distance arborée

- Dans un arbre, la distance δ_{ij} entre deux UTO i et j est donnée par la somme des longueurs de branches les séparant :
  - On parle de distance *arborée* ou *patristique* :
    - Doit vérifier, en plus des trois conditions standard, la *condition des quatre points* ($\delta_{ij} + \delta_{kl} \leq \max(\delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{jk})$).
- Objectif des différentes méthodes de distances :
  - Faire que les valeurs δ_{ij} correspondent le plus fidèlement possible aux valeurs de d_{ij} présentes dans la matrice de départ.



Typologie

- Méthodes nécessitant d'explorer l'ensemble des topologies (optimisation d'un critère) :
  - Moindres carrés (*Least Squares*, LS) :
  - Mimum d'évolution (*Minimum of Evolution*, ME).
- Méthodes construisant un arbre unique :
  - Classification ascendante hiérarchique au lien moyen (*Unweighted Pair-Group Method with Arithmetic means*, UPGMA).
  - *Neighbor Joining* (NJ).

Principe général

- Pour une topologie τ donnée, déterminer quelles sont les valeurs des longueurs de branches minimisant :

$$Q = \sum_{i<j} w_{ij}(d_{ij} - \delta_{ij})^2$$

avec w_{ij} les valeurs de pondération associées à chaque paire (i, j) :

  - Pondération uniforme ($w_{ij} = 1$).
  - Inverse de la distance ($w_{ij} = 1/d_{ij}$).
  - Inverse du carré de la distance ($w_{ij} = 1/d_{ij}^2$).

- Effectuer ces calculs pour l'ensemble des topologies possibles :
  - Retenir celle pour laquelle Q est minimale.

Moindres carrés standard

- Soit b_k la longueur de la branche k de l'arbre à n UTO considéré ($1 \leq k \leq 2n-3$).
- Soit $x_{ij,k}$ une variable indicatrice telle que :
  - $x_{ij,k} = 1$ si la branche k se situe sur le chemin allant du taxon i au taxon j.
  - $x_{ij,k} = 0$ dans le cas contraire.
- Dans ce cas, la valeur de la distance patristique entre i et j est égale à $\delta_{ij} = \sum_k x_{ij,k}\, b_k$, et Q peut s'écrire comme :

$$Q = \sum_{i<j} w_{ij} \left(d_{ij} - \sum_{k=1}^{2n-3} x_{ij,k}\, b_k \right)^2$$

Expression matricielle

- Soient $\mathbf{b}$, $\mathbf{d}$, $\mathbf{X}$ et $\mathbf{W}$ tels que :
  - $\mathbf{b} = (b_1, b_2, \ldots, b_k)$, le vecteur des longueurs de branches.
  - $\mathbf{d} = (d_{12}, d_{13}, \ldots, d_{n-1\,n})$, le vecteur ordonné contenant l'ensemble des distances.
  - $\mathbf{X} = (x_{ij,k})$, la matrice des valeurs de $x_{ij,k}$ ordonnées de façon à ce que chaque ligne de $\mathbf{X}$ corresponde aux lignes de $\mathbf{d}$.
  - $\mathbf{W} = \mathrm{diag}(w_{12}, w_{13}, \ldots, w_{n-1\,n})$, la matrice diagonale des pondérations.
- Dans ce cas, l'expression matricielle permettant de déterminer $\mathbf{b}$ de façon à minimiser Q est :

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{d}$$

soit la solution standard du problème des moindres carrés.

Jeu de données exemple

- Jeu de données de Brown *et al.* (1982) sur les séquences d'ADN mitochondrial d'Hominoïdes.
- Modèle de Kimura à deux paramètres pour le calcul de la matrice de distances :

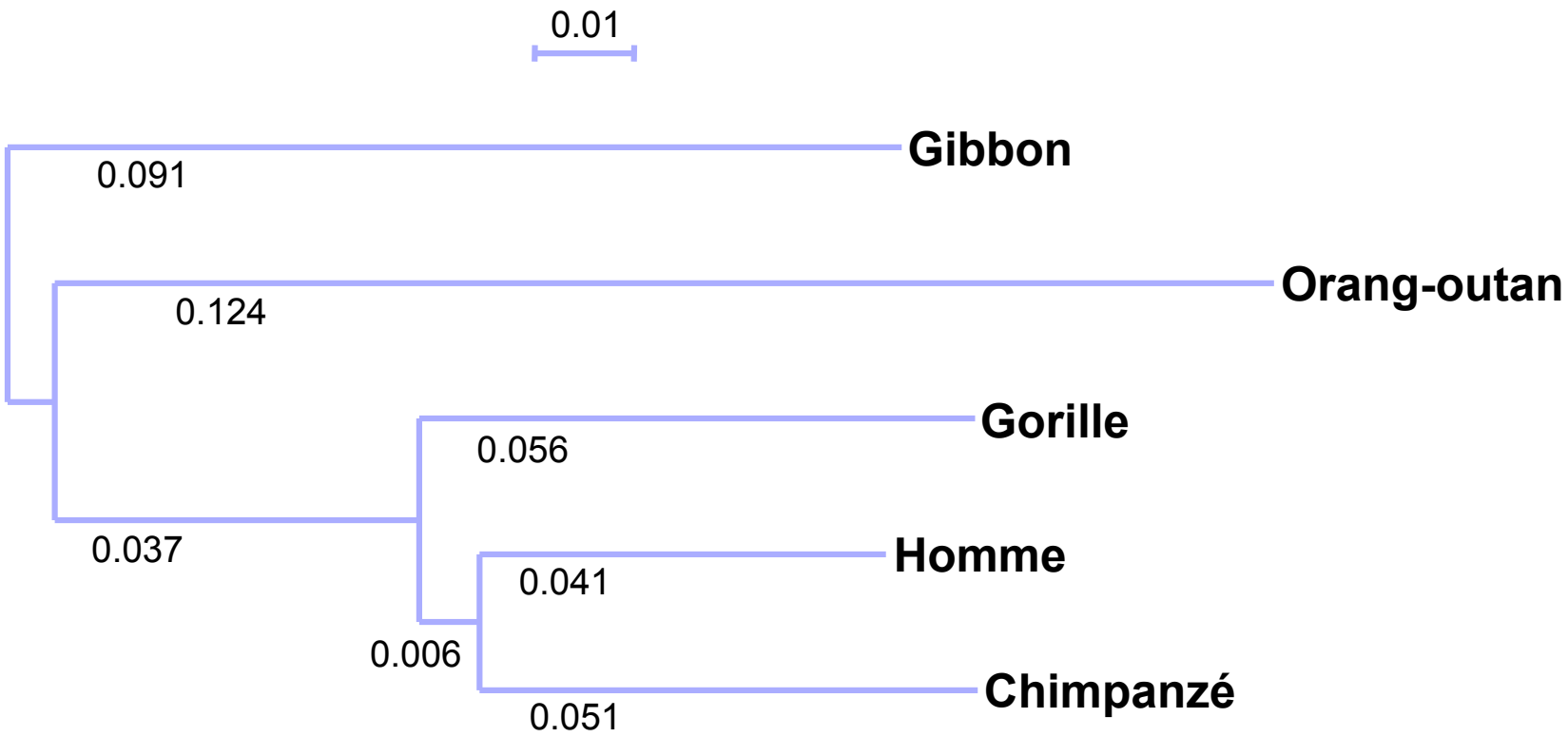
$\mathbf{D} = (d_{ij})$

	H	C	G	O	B
H	0				
C	0.092	0			
G	0.106	0.111	0		
O	0.177	0.193	0.188	0	
B	0.207	0.218	0.218	0.219	0

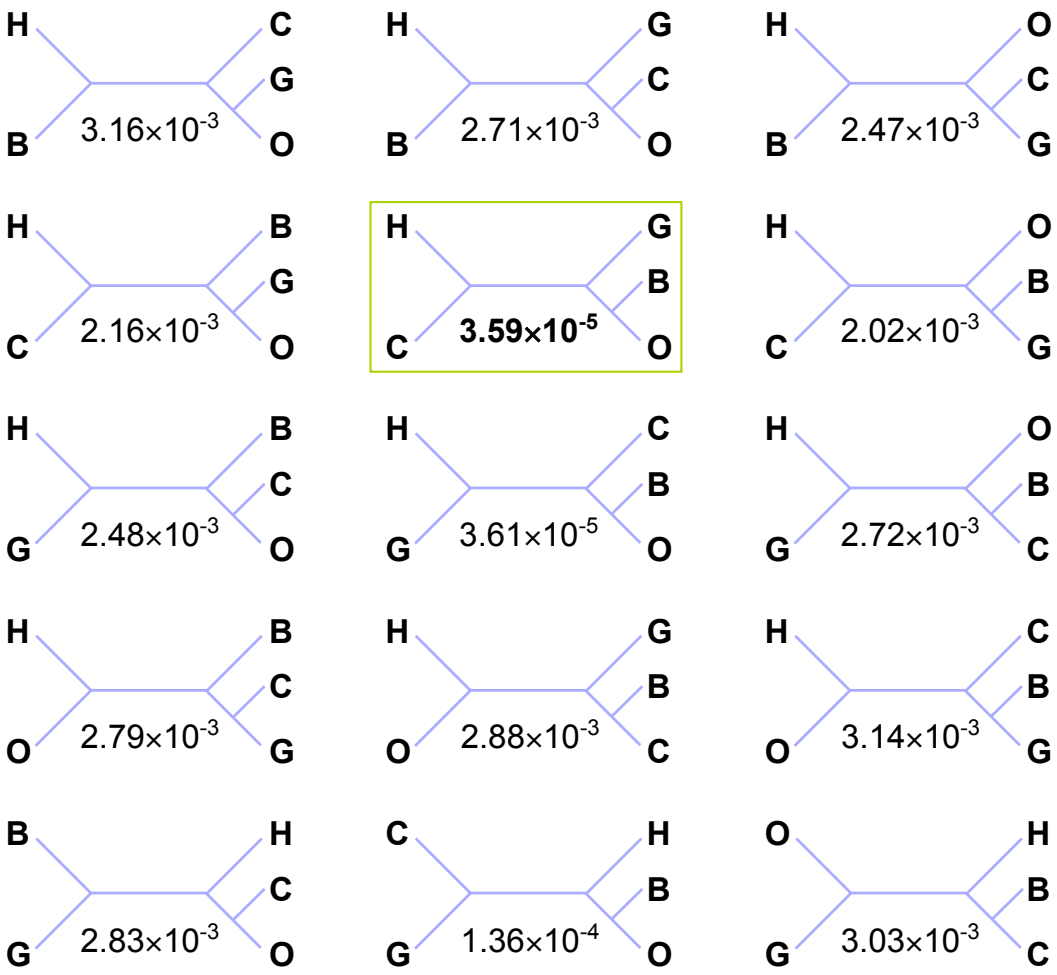
Humain = H
Chimpanzé = C
Gorille = G
Orang-outan = O
Gibbon = B

Arbre obtenu

- Pondération par l'inverse du carré de la distance.
- Racinement par la séquence du Gibbon.



Scores des topologies



B = Gibbon, H = Homme, C = Chimpanzé, G = Gorille, O = Orang-outan

Commentaires sur le résultat

- La différence de scores entre la bonne topologie (retenue) et la deuxième meilleure porte sur la septième décimale :
  - Quelle est la significativité de cette différence ?
- L'utilisation de la pondération uniforme ou par l'inverse de la distance ne permettent pas de retenir la topologie vraie.

Avantages et limitations

- Méthode consistante.
- Algorithme de complexité en $O(n^3)$:
  - Inversion de $\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}$.
- Aussi efficace que le maximum de vraisemblance si les variables suivent une distribution normale :
  - Nécessité d'avoir un grand nombre de sites dans l'alignement.
- Peut donner des longueurs de branches négatives.
- Problèmes de dérives numériques si la matrice est mal conditionnée (*i.e.*, $\det(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}) \simeq 0$) :
  - Utilisation de simplifications ne nécessitant pas d'effectuer une inversion de matrice :
    - Approximation de Fitch et Margoliash (1967).
    - Simplification de Rzhetsky et Nei (1992).

Approximation de Fitch et Margoliash

- Estimations moins précises que celles obtenues par les moindres carrés proprement dits :
  - Différences observées souvent négligeables.
- Construction en effectuant des groupements par triplets :
  - Correspondance exacte entre distance observée et la distance patristique :
    - Calcul simple des longueurs de branches.
  - Soit d_{AB}, d_{AC} et d_{BC} les valeurs des distances entre trois groupes A, B et C, dans ce cas, il est possible d'écrire que :

$$\left\{\begin{array}{l} d_{AB} = b_A + b_B \\ d_{AC} = b_A + b_C \\ d_{BC} = b_B + b_C \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} b_A = (d_{AB} + d_{AC} - d_{BC})/2 \\ b_B = (d_{AB} + d_{BC} - d_{AC})/2 \\ b_C = (d_{AC} + d_{BC} - d_{AB})/2 \end{array}\right.$$

Algorithme I

Pour chacune des $n(n-1)/2$ paires (i, j) possibles, faire :

1. $A \leftarrow i$, $B \leftarrow j$ et regroupement de toutes les autres UTO dans C.
2. Calcul des distances d_{AC} et d_{BC} telles que :

$$d_{AC} = \frac{1}{n_C} \sum_{j \in C} d_{Aj} \quad \text{et} \quad d_{BC} = \frac{1}{n_C} \sum_{j \in C} d_{Bj}$$

   avec $n_C = \text{card}(C)$ le nombre d'éléments présents dans C.
3. Calcul des trois longueurs de branches au moyen de la formule précédente :
   - Soustraction des longueurs déjà calculées le cas échéant.

Algorithme II

4. Regrouper A et B dans un même ensemble $Z = A \cup B$ puis calculer, pour chaque $j \in C$:

$$d_{Zj} = \frac{1}{n_Z} \sum_{i \in Z} d_{ij}$$

avec $n_Z = \text{card}(Z)$, le nombre d'éléments présents dans Z. Les valeurs obtenues remplacent celles correspondant à A et à B.

5. Si $\dim(\mathbf{D}) \geq 3$, alors :
   - Réinitialiser A et B avec les UTO ou les groupes d'UTO pour lesquels d_{ij} est minimale et retourner en 2.

   Sinon, aller en 6.

6. Calcul de la valeur de Q.

Exemple d'utilisation I

- Initialisation en prenant la paire (i, j) telle que d_{ij} soit minimale :

$$A \leftarrow \{\mathrm{H}\},\ B \leftarrow \{\mathrm{C}\} \text{ et } C \leftarrow \{\mathrm{G}, \mathrm{O}, \mathrm{B}\}$$

- Calcul de d_{AB}, d_{AC} et d_{BC} :

$$\begin{cases} d_{AB} = 0.092 \\ d_{AC} = (0.106 + 0.177 + 0.207)/3 = 0.163 \\ d_{BC} = (0.111 + 0.193 + 0.218)/3 = 0.174 \end{cases}$$

- Calcul des longueurs de branches correspondantes :

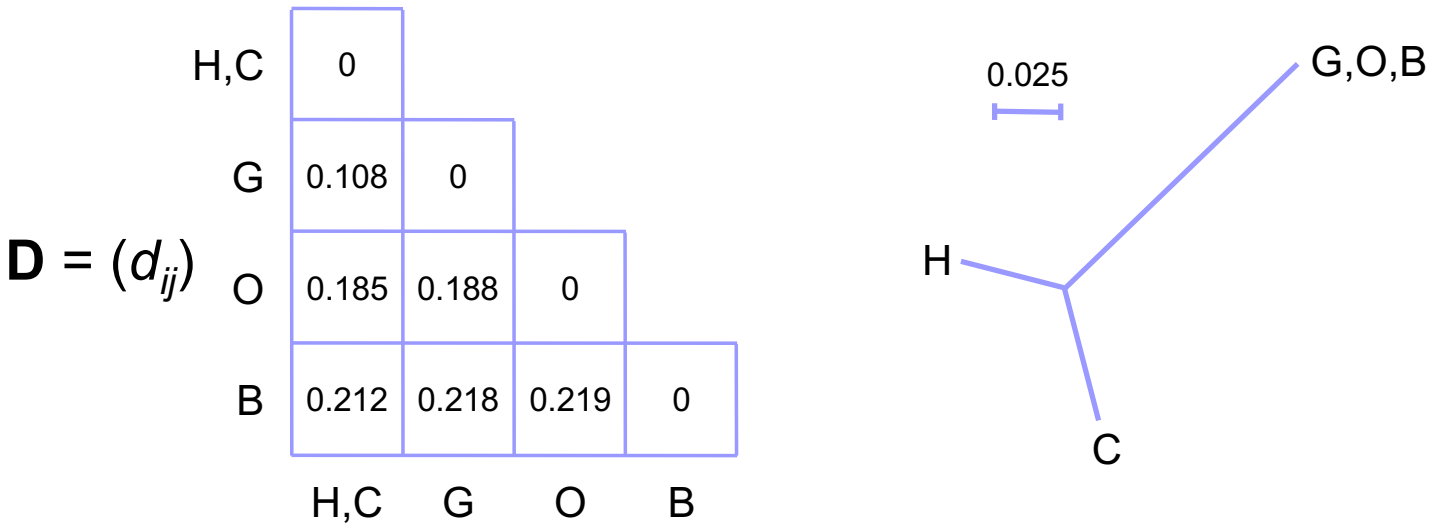
$$\begin{cases} b_A = (0.092 + 0.163 - 0.174)/2 = 0.041 \\ b_B = (0.092 + 0.174 - 0.163)/2 = 0.052 \\ b_C = (0.163 + 0.174 - 0.092)/2 = 0.123 \end{cases}$$

Exemple d'utilisation II

- Calcul des nouvelles distances, avec $Z = A \cup B = \{\mathrm{H}, \mathrm{C}\}$:

$$\begin{cases} d_{Z\mathrm{G}} = (0.106 + 0.111)/2 = 0.108 \\ d_{Z\mathrm{O}} = (0.177 + 0.193)/2 = 0.185 \\ d_{Z\mathrm{B}} = (0.207 + 0.218)/2 = 0.212 \end{cases}$$

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



Exemple d'utilisation III

- Du fait que $\dim(\mathbf{D}) \geq 3$, on relance une itération avec :

$$A \leftarrow \{\mathrm{H}, \mathrm{C}\},\ B \leftarrow \{\mathrm{G}\} \text{ et } C \leftarrow \{\mathrm{O}, \mathrm{B}\}$$

- Calcul de d_{AB}, d_{AC} et d_{BC} :

$$\begin{cases} d_{AB} = 0.108 \\ d_{AC} = (0.185 + 0.212)/2 = 0.199 \\ d_{BC} = (0.188 + 0.218)/2 = 0.203 \end{cases}$$

- Calcul des longueurs de branches correspondantes :

$$\begin{cases} b_A = (0.108 + 0.199 - 0.203)/2 = 0.052 \\ b_B = (0.108 + 0.203 - 0.199)/2 = 0.056 \\ b_C = (0.199 + 0.203 - 0.108)/2 = 0.147 \end{cases}$$

Exemple d'utilisation IV

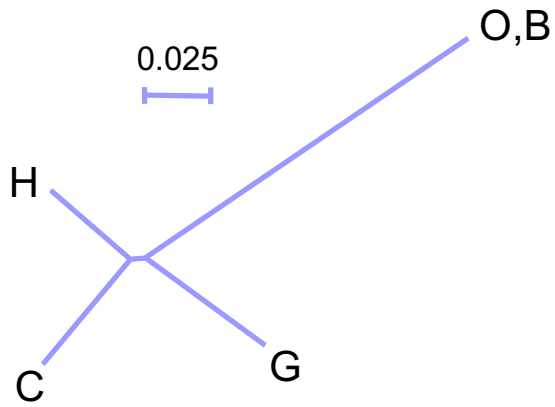
- Dans le cas de b_A, prise en compte des longueurs de branches existantes conduisant aux éléments de A :
  - La longueur de la branche interne à ajouter est égale à $0.052 - (0.0405 + 0.0515)/2 = 0.006$.
- Calcul des nouvelles distances, avec $Z = A \cup B = \{\{\mathrm{H}, \mathrm{C}\}, \mathrm{G}\}$:

$$\begin{cases} d_{Z\mathrm{O}} = (0.185 + 0.188)/2 = 0.186 \\ d_{Z\mathrm{B}} = (0.212 + 0.218)/2 = 0.215 \end{cases}$$

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenus :

$\mathbf{D} = (d_{ij})$

H,C,G	0		
O	0.186	0	
B	0.215	0.219	0
	H,C,G	O	B



Exemple d'utilisation V

- Dernière itération avec :

$$A \leftarrow \{\mathrm{H}, \mathrm{C}, \mathrm{G}\},\ B \leftarrow \{\mathrm{O}\} \text{ et } C \leftarrow \{\mathrm{B}\}$$

- Calcul de d_{AB}, d_{AC} et d_{BC} :

$$\begin{cases} d_{AB} = 0.186 \\ d_{AC} = 0.215 \\ d_{BC} = 0.219 \end{cases}$$

- Calcul des longueurs de branches correspondantes :

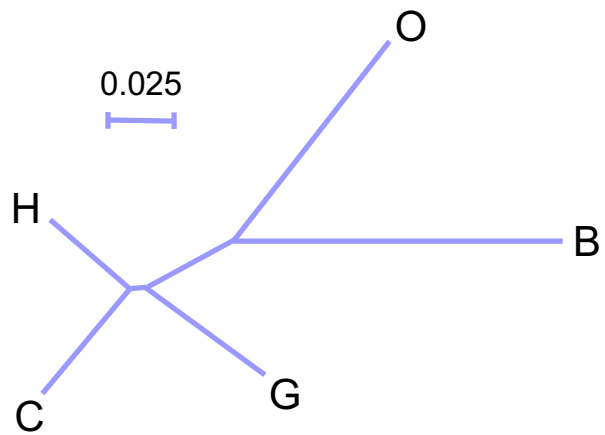
$$\begin{cases} b_A = (0.186 + 0.215 - 0.219)/2 = 0.091 \\ b_B = (0.186 + 0.219 - 0.215)/2 = 0.095 \\ b_C = (0.215 + 0.219 - 0.186)/2 = 0.124 \end{cases}$$

Exemple d'utilisation VI

- Dans le cas de b_A prise en compte des longueurs de branches existantes conduisant aux éléments de A :
  - La longueur de la branche interne à ajouter est égale à $0.091 - (0.0405 + 0.006 + 0.0515 + 0.006 + 0.056)/3 = 0.038$.
- Matrice des distances patristiques et arbre obtenus :

$\Delta = (\delta_{ij})$

	H	C	G	O	B
H	0				
C	0.092	0			
G	0.103	0.113	0		
O	0.179	0.191	0.189	0	
B	0.208	0.220	0.218	0.219	0



Avantages et limitations

- Calcul simultané de la topologie et des longueurs de branches.
- Pas d'exploration de l'ensemble des topologies :
  - Seulement $n(n-1)/2$ itérations (*i.e.*, le nombre de paires possibles entre deux UTO) :
    - Complexité globale de l'algorithme en $O(n^5)$.
  - Pas de garantie que l'arbre obtenu soit effectivement celui des moindres carrés.

Minimum d'évolution

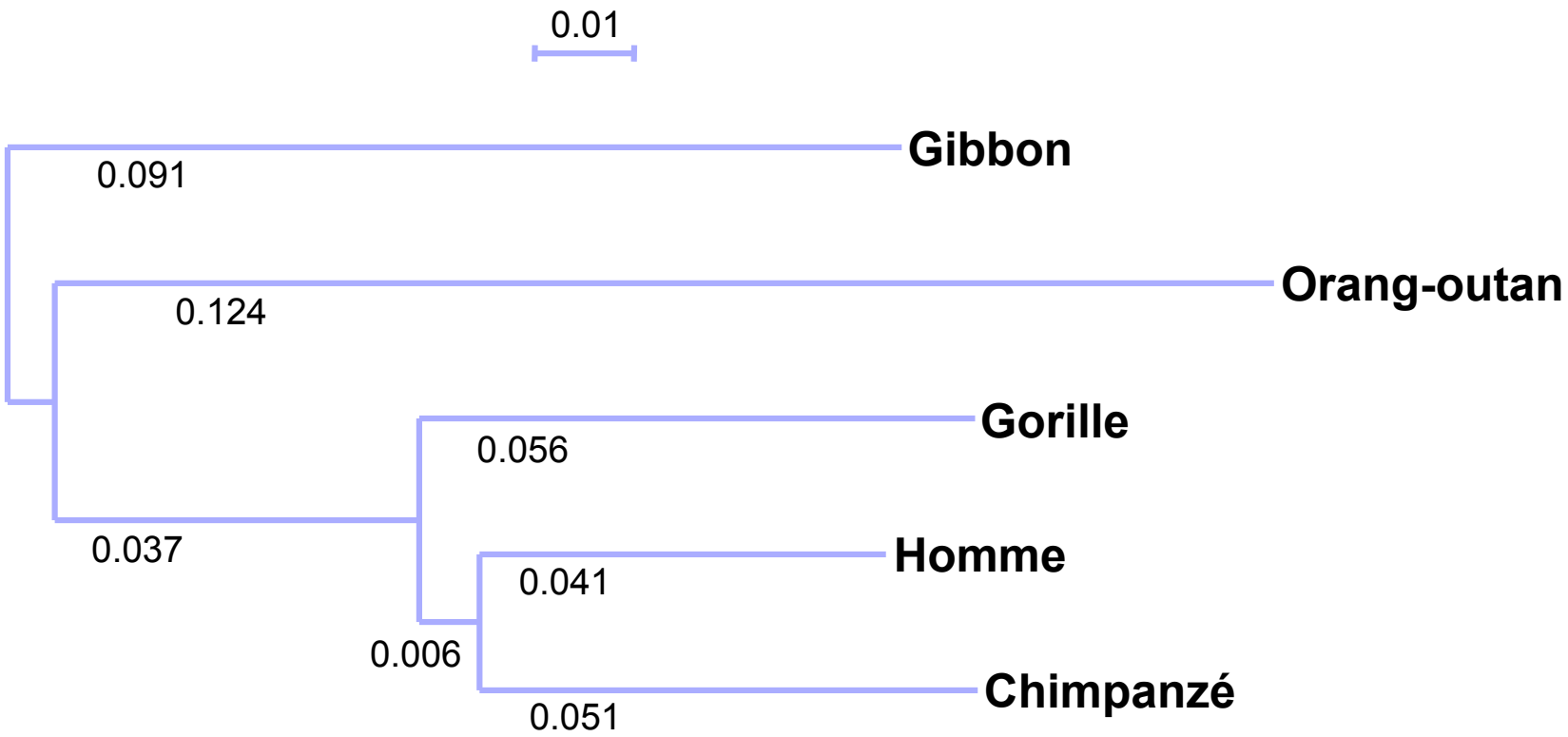
- Méthode très comparable aux moindres carrés (mêmes avantages et mêmes inconvénients).
- Pour une topologie τ donnée :
  - Détermination des longueurs de branches par les moindres carrés.
  - Calcul de la longueur de l'arbre S, telle que :

$$S = \sum_{k=1}^{2n-3} b_k$$

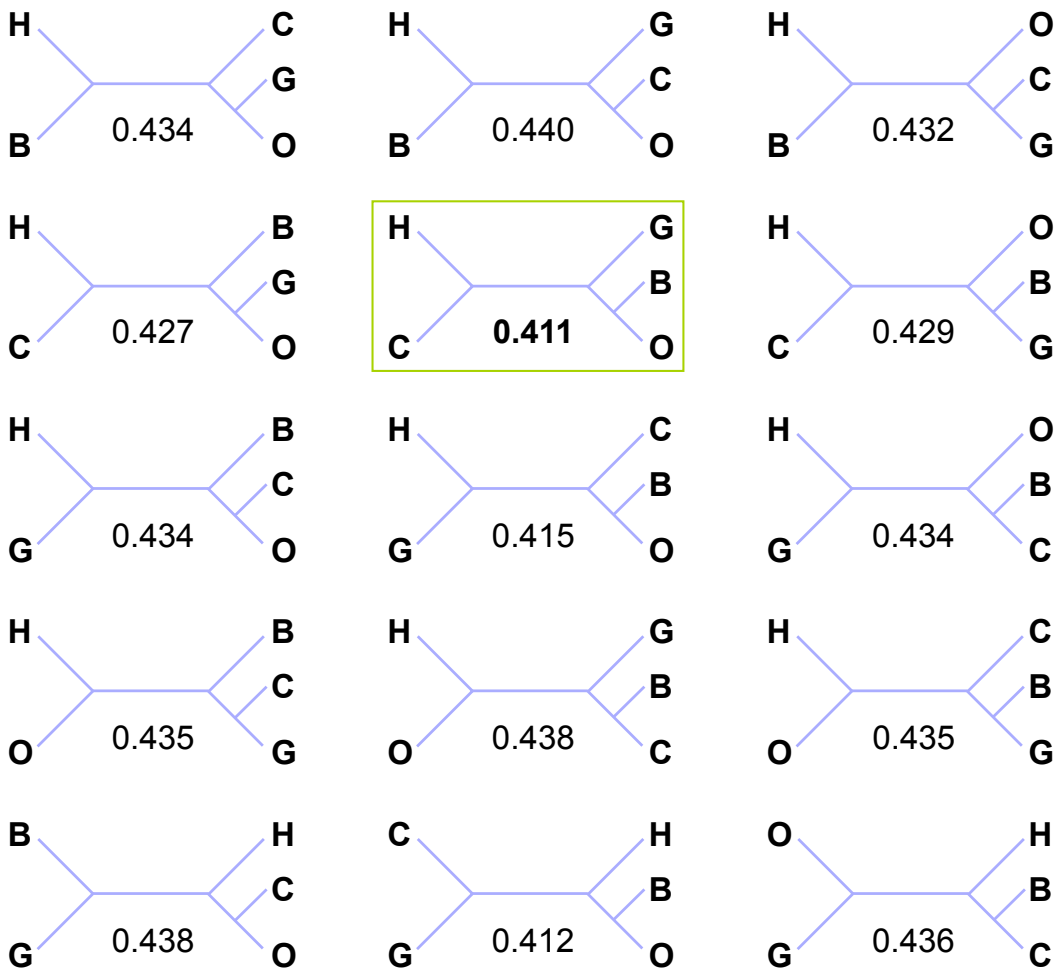
- Effectuer ces calculs pour l'ensemble des topologies possibles :
  - Retenir celle pour laquelle S est minimale.

Arbre obtenu

- Mêmes paramamètres que pour les moindres carrés.
- Même topologie retenue, et donc mêmes longueurs de branches.
- Racinement par la séquence du Gibbon.



Scores des topologies



B = Gibbon, H = Homme, C = Chimpanzé, G = Gorille, O = Orang-outan

Classification ascendante hiérarchique

- Méthode la plus simple du point de vue algorithmique.
- Tire son nom du fait que la construction de l'arbre démarre à partir des feuilles.
- Une des seules à produire des arbres enracinés.
- Les distances patristiques générées par cette méthode sont dites *ultramétriques* :
  - Doivent satisfaire la condition dite *d'inégalité ultratriangulaire* ($\delta_{ik} \leq \max(\delta_{ij}, \delta_{jk})\ \forall i, j, k$).
  - Les longueurs des chemins allant de la racine à n'importe quelle feuille sont égales.

Algorithme I

Tant que $\dim(\mathbf{D}) > 1$:

1. Identifier les deux ensembles d'UTO C_i et C_j pour lesquels d_{ij} est minimale.
2. Créer l'ensemble C_u tel que $C_u \leftarrow C_i \cup C_j$, avec u une UTH nouvellement créée.
3. Connecter C_i et C_j à u et attribuer aux deux branches reliant u à C_i et C_j la longueur $d_{ij}/2$:
   - Tout comme dans le cas de Fitch et Margoliash, soustraction éventuelle des longueurs déjà calculées pour les branches internes.

Algorithme II

5. Calculer la distance entre C_u et chacun des k autres groupes présents dans $\mathbf{D}$ (exceptés C_i et C_j) au moyen de :

$$d_{uk} = \frac{n_i}{n_i + n_j} d_{ik} + \frac{n_j}{n_i + n_j} d_{jk}$$

avec $n_i = \text{card}(C_i)$ et $n_j = \text{card}(C_j)$.

6. Supprimer de $\mathbf{D}$ les lignes et colonnes correspondant à C_i et C_j et ajouter la ligne et la colonne correspondant à C_u avec les valeurs de d_{uk}.

Exemple d'utilisation I

- Initialisation avec $C_i \leftarrow \{\mathrm{H}\}$ et $C_j \leftarrow \{\mathrm{C}\}$.
- $C_u \leftarrow C_i \cup C_j = \{\mathrm{H}, \mathrm{C}\}$
- Calcul des longueurs de branches conduisant à u :

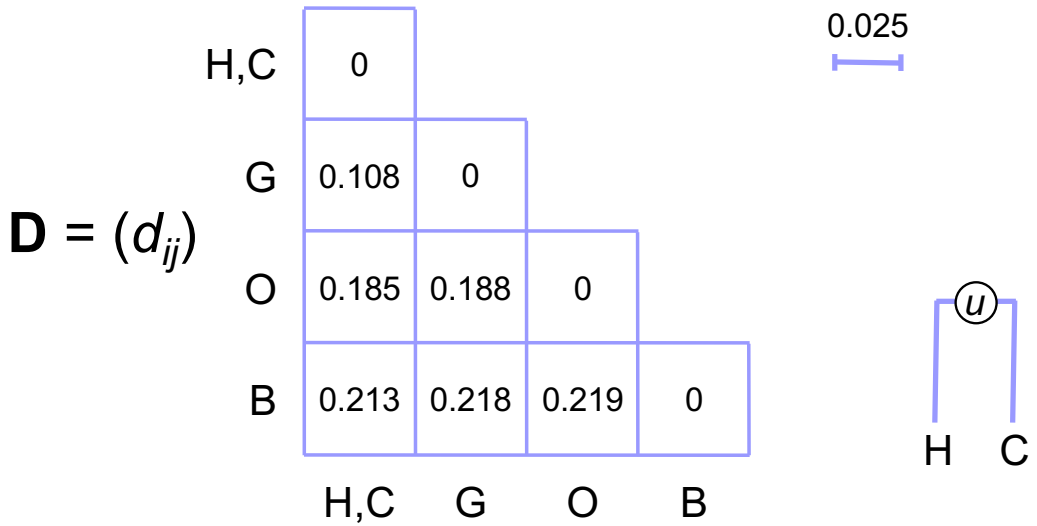
$$b_{ui} = b_{uj} = 0.092/2 = 0.046$$

- Calcul des distances entre C_u et les trois autres groupes présents dans $\mathbf{D}$ (*i.e.*, {G}, {O} et {B}) :

$$\begin{cases} d_{u\mathrm{G}} = 0.106/2 + 0.111/2 = 0.108 \\ d_{u\mathrm{O}} = 0.177/2 + 0.193/2 = 0.185 \\ d_{u\mathrm{B}} = 0.207/2 + 0.218/2 = 0.213 \end{cases}$$

Exemple d'utilisation II

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



- Démarrage de la $2^{\text{ème}}$ itération avec $C_i \leftarrow \{\text{H}, \text{C}\}$ et $C_j \leftarrow \{\text{G}\}$.
- $C_u \leftarrow C_i \cup C_j = \{\text{H}, \text{C}, \text{G}\}$

Exemple d'utilisation III

- Calcul des longueurs de branches conduisant à u :

$$b_{ui} = b_{uj} = 0.108/2 = 0.054$$

Dans le cas de i, il existe déjà une branche de longueur 0.046 reliant {H} à son ancêtre commun avec {C} :

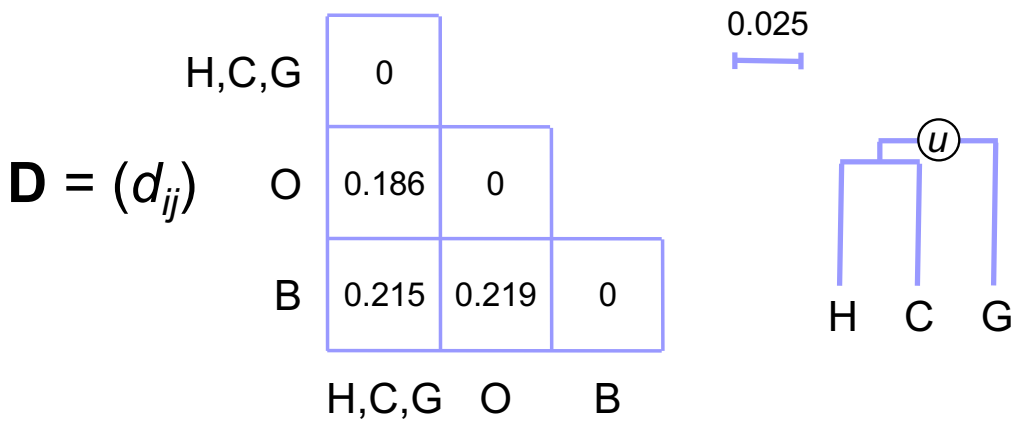
  - La longueur de la branche interne à ajouter est égale à $0.054 - 0.046 = 0.008$.

- Calcul des distances entre C_u et les deux autres groupes présents dans $\mathbf{D}$ (*i.e.*, {O} et {B}) :

$$\begin{cases} d_{uO} = 2/3 \times 0.185 + 1/3 \times 0.188 = 0.186 \\ d_{uB} = 2/3 \times 0.213 + 1/3 \times 0.218 = 0.215 \end{cases}$$

Exemple d'utilisation IV

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



- Démarrage de la $3^{\text{ème}}$ itération avec $C_i \leftarrow \{\text{H}, \text{C}, \text{G}\}$ et $C_j \leftarrow \{\text{O}\}$.
- $C_u \leftarrow C_i \cup C_j = \{\text{H}, \text{C}, \text{G}, \text{O}\}$

Exemple d'utilisation V

- Calcul des longueurs de branches conduisant à u :

$$b_{ui} = b_{uj} = 0.186/2 = 0.093$$

Dans le cas de i, prise en compte des longueurs de branches conduisant aux feuilles :

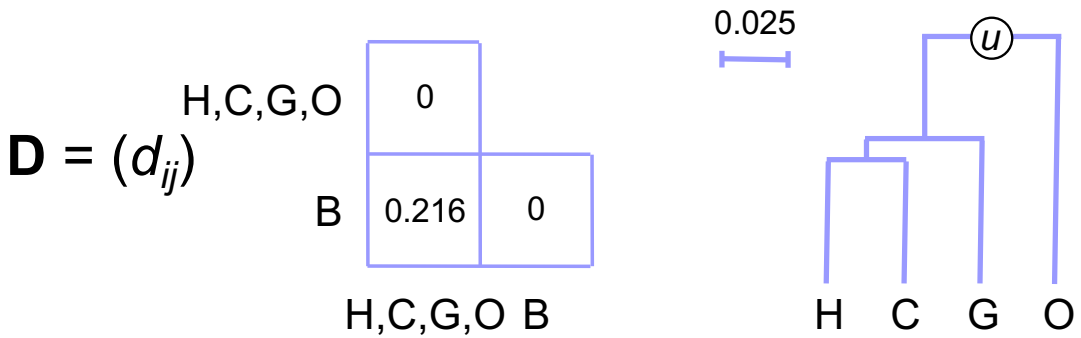
  - La longueur de la branche interne à ajouter est égale à $0.093 - 0.046 - 0.008 = 0.039$.

- Calcul des distances entre C_u et le dernier groupe présent dans $\mathbf{D}$ (*i.e.*, $\{\mathrm{B}\}$) :

$$d_{u\mathrm{B}} = 3/4 \times 0.215 + 1/4 \times 0.219 = 0.216$$

Exemple d'utilisation VI

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :



- Démarrage de la dernière itération avec $C_i \leftarrow \{\mathrm{H}, \mathrm{C}, \mathrm{G}, \mathrm{O}\}$ et $C_j \leftarrow \{\mathrm{B}\}$.
- $C_u \leftarrow C_i \cup C_j = \{\mathrm{H}, \mathrm{C}, \mathrm{G}, \mathrm{O}, \mathrm{B}\}$

Exemple d'utilisation VII

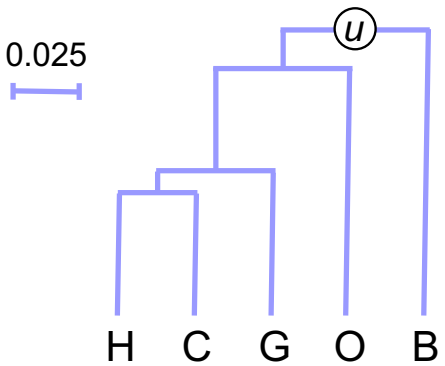
- Calcul des longueurs de branches conduisant à u :

$$b_{ui} = b_{uj} = 0.216/2 = 0.108$$

Dans le cas de i, prise en compte des longueurs de branches conduisant aux feuilles :

- La longueur de la branche interne à ajouter est égale à $0.108 - 0.046 - 0.008 - 0.039 = 0.015$.

- Arbre raciné final :



Avantages et limitations

- Complexité en $O(n^3)$, ce qui en fait une méthode très rapide, utilisable même avec des milliers d'UTO.
- Valide uniquement dans le cas où les vitesses d'évolution sont les mêmes dans toutes les lignées :
  - Hypothèse de *l'horloge moléculaire*.
  - Utilisation limitée à des séquences proches du point de vue évolutif.
- N'est plus employée en phylogénie.
- Est encore utilisée pour des problèmes de classification nécessitant de travailler sur des matrices de distances de grande taille.

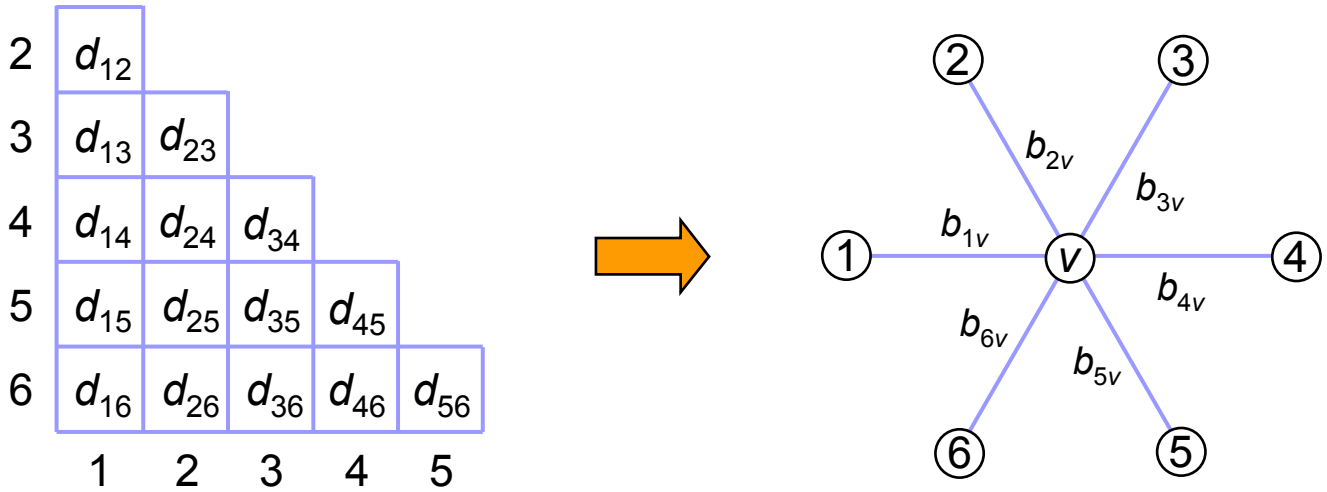
Algorithme I

1. Initialisation à partir d'une topologie en étoile telle que :
   - Branches b_{iv} reliées à un nœud central v.
   - Expression des valeurs de d_{ij} à partir des longueurs de branches :

$$d_{ij} = b_{iv} + b_{jv} \ (i \neq j)$$

   - Longueur de l'arbre déduite :

$$S_0 = \sum_{i=1}^{n} b_{iv} = \frac{1}{n-1} \sum_{i<j} d_{ij}$$

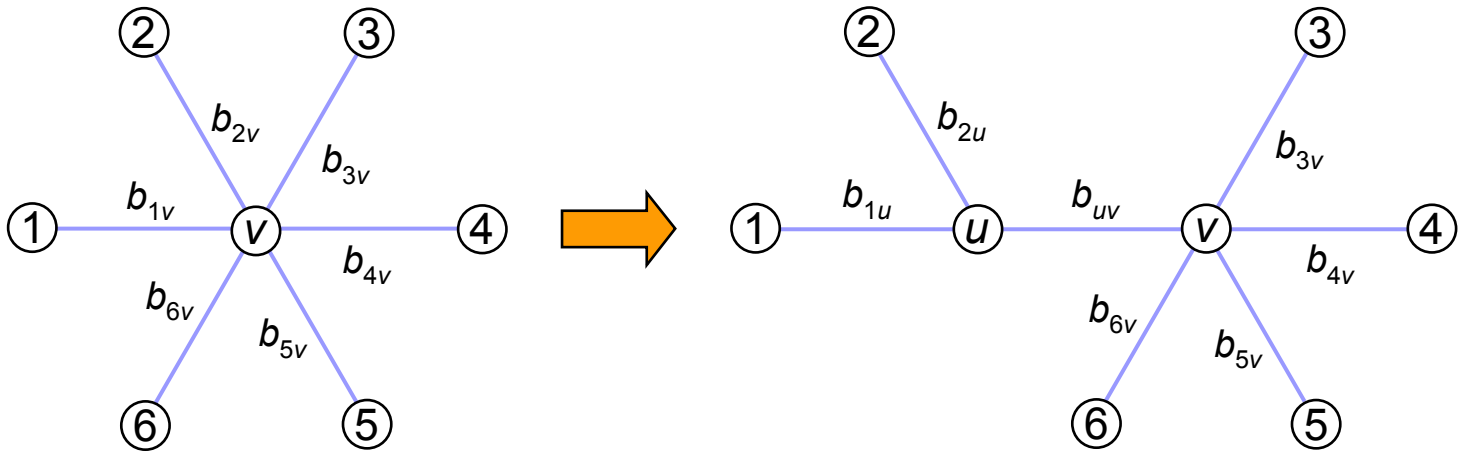


Algorithme II

2. Identification de la paire (i, j) qui, une fois agglomérée, minimise la longueur de l'arbre S_{ij} :
   - Création d'un nœud u connectant i et j.
   - Création d'une branche interne b_{uv} connectant u et v.
   - Dans ce cas, expression de S_{ij} comme :

$$\begin{aligned} S_{ij} &= b_{iu} + b_{ju} + b_{uv} + S_k \\ &= d_{ij} + b_{uv} + S_k \end{aligned}$$

avec S_k la longueur de l'arbre en étoile contenant les $n - 2$ UTO restantes.



Algorithme III

3. Sachant que :

$$S_k = \sum_{k \neq i,j} b_{kv} = \frac{1}{n-3} \sum_{k \neq i,j; k<l} d_{kl}$$

et que :

$$b_{uv} = \frac{1}{2(n-2)} \left[\sum_{k \neq i,j} (d_{ik} + d_{jk}) - (n-2)\, d_{ij} - 2S_k \right]$$

on déduit l'expression de S_{ij} :

$$S_{ij} = \frac{1}{2} d_{ij} + \frac{1}{2(n-2)} \sum_{k \neq i,j} (d_{ik} + d_{jk}) + \frac{1}{n-2} \sum_{k \neq i,j; k<l} d_{kl}$$

Algorithme IV

4. Une fois la paire (i, j) identifiée, recalcul des longueurs de branches b_{iu} et b_{ju} au moyen de Fitch-Margoliash :

$$b_{iu} = \frac{1}{2}\left(d_{ij} + \frac{1}{n-2}\sum_{k\neq i,j} d_{ik} - \frac{1}{n-2}\sum_{k\neq i,j} d_{jk}\right)$$

et :

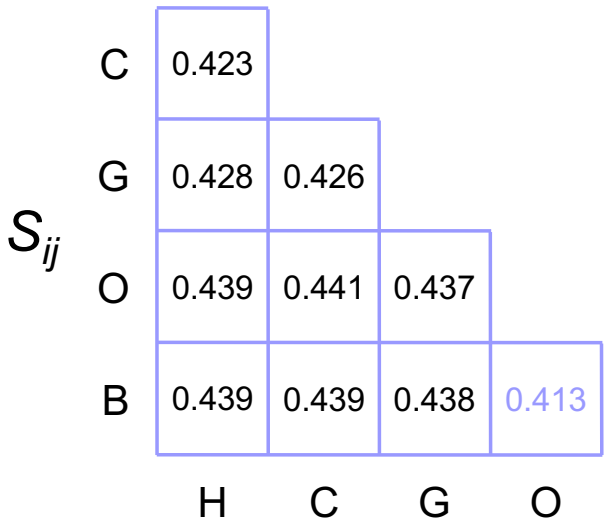
$$b_{ju} = \frac{1}{2}\left(d_{ij} + \frac{1}{n-2}\sum_{k\neq i,j} d_{jk} - \frac{1}{n-2}\sum_{k\neq i,j} d_{ik}\right)$$

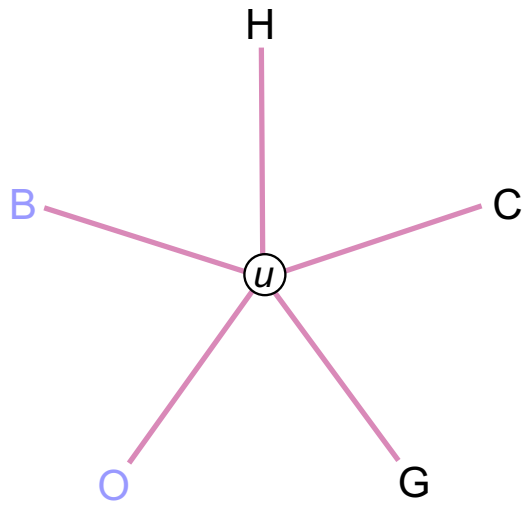
5. Recalcul de la matrice $\mathbf{D}$ en remplaçant les lignes correspondant à i et j par la paire (i, j), telle que :

$$d_{ij,k} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$$

Exemple d'utilisation I

- Initialisation à partir d'une topologie en étoile de longueur $S_0 = 0.432$.
- Calcul de l'ensemble des valeurs de S_{ij} possibles :
  - Identification de la paire (O, B) comme étant celle minimisant S_{ij} :





Exemple d'utilisation II

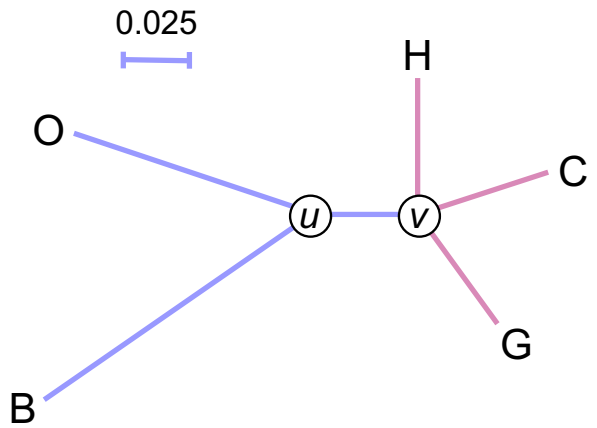
- Calcul des longueurs de branches conduisant à u et calcul de la longueur de la branche interne b_{uv} :

$$b_{\mathrm{O}u} = 0.0955,\ b_{\mathrm{B}u} = 0.1238 \text{ et } b_{uv} = 0.0392$$

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :

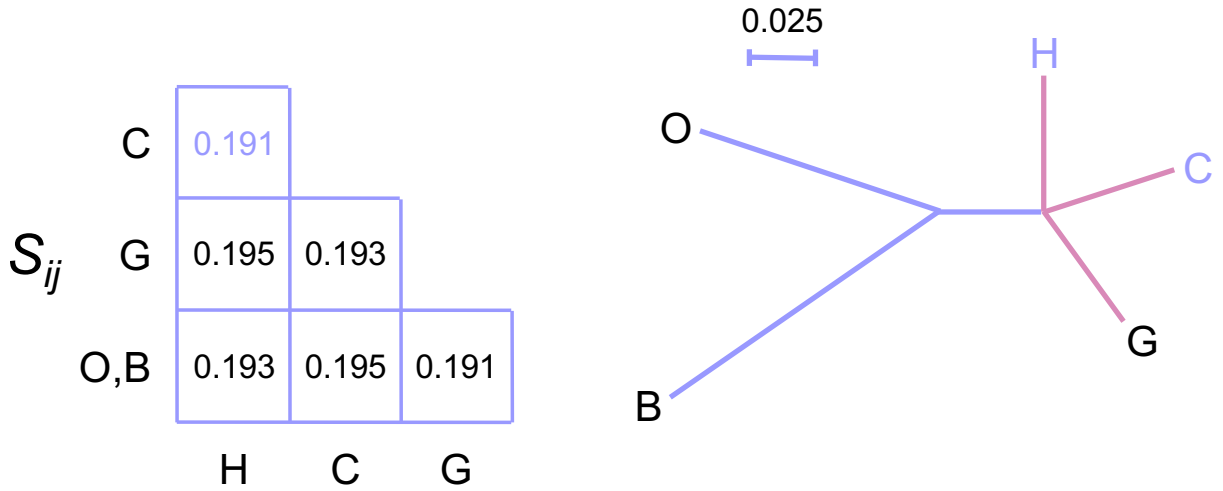
$\mathbf{D} = (d_{ij})$

	H	C	G	O,B
H	0			
C	0.092	0		
G	0.106	0.111	0	
O,B	0.082	0.096	0.094	0



Exemple d'utilisation III

- Calcul de l'ensemble des nouvelles valeurs de S_{ij} possibles :
  - Identification de la paire (H, C) comme étant celle minimisant S_{ij} :



- Calcul des longueurs de branches conduisant à u et calcul de la longueur de la branche interne b_{uv} :

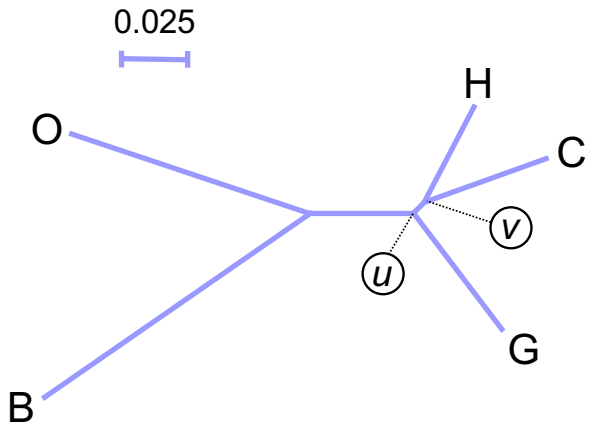
$$b_{\mathrm{H}u} = 0.0413, b_{\mathrm{C}u} = 0.0505 \text{ et } b_{uv} = 0.006$$

Exemple d'utilisation IV

- Nouvelles valeurs de $\mathbf{D}$ et arbre obtenu :

$\mathbf{D} = (d_{ij})$

	H,C	G	O,B
H,C	0		
G	0.062	0	
O,B	0.043	0.094	0



- Calcul de la longueur de la branche conduisant à {G} en utilisant Fitch-Margoliash, soit :

$$(0.0623 + 0.0936 - 0.0432)/2 = 0.0564$$

Avantages et limitations

- Méthode consistante.
- À chaque itération, les longueurs de branches calculées sont une estimation de celles obtenues aux moindres carrés.
- Rapide, même avec des milliers d'UTO :
  - Implémentation originale par Saitou et Nei (1987) avec une complexité en $O(n^5)$.
  - Amélioration de Studier et Keppler (1988) réduisant la complexité en $O(n^3)$.
  - Dernière amélioration par Gascuel (1997) minimisant la variance de $\mathbf{D}$ à chaque recalcul de la matrice.
- L'arbre obtenu est une bonne approximation de l'arbre du minimum d'évolution.

Plan

Maximum de vraisemblance

- 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 désignée par le terme $S^{(i)}$ $(1 \leq i \leq \ell)$.
- Par ailleurs, le vecteur des paramètres est $\boldsymbol{\theta} = (\tau, \mathbf{t}, \boldsymbol{\vartheta}, \alpha)$, avec :
  - τ la topologie de l'arbre.
  - $\mathbf{t}$ le vecteur des longueurs de branches.
  - $\boldsymbol{\vartheta}$ le vecteur des paramètres 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 $\boldsymbol{\theta}$:

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

Modèle de Jukes et Cantor

- Le calcul de la divergence observée entre deux séquences au moyen du modèle de Jukes et Cantor est donnée par (*cf.* Diapo. 56) :

$$p = 3p_{ij}(t) = \frac{3}{4} - \frac{3}{4}\mathrm{e}^{-4t/3} \quad (i \neq j)$$

- 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 t est donnée par la loi binomiale $\mathcal{B}(\ell, p)$ telle que :

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

Simplification des calculs

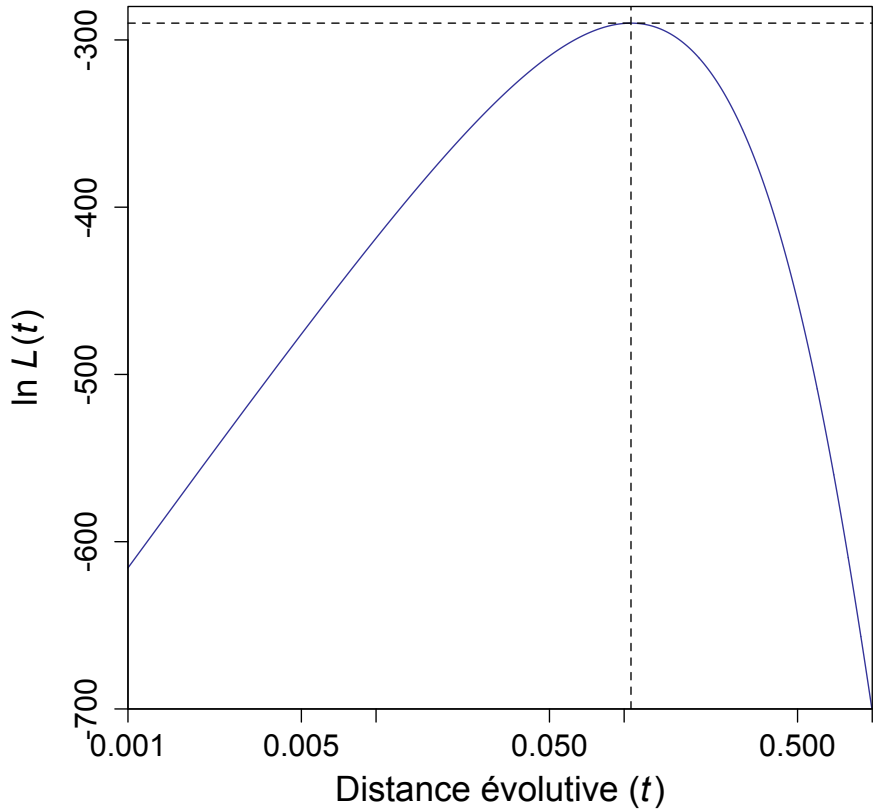
- Le coefficient binomial $\binom{\ell}{n}$ étant une constante, il peut être omis pour effectuer les calculs :
  - La vraisemblance obtenue change, mais le maximum sera toujours obtenu pour la même valeur de t.
- Passage en logarithmes pour éviter les dérives numériques du fait que les valeurs attendues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

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

- 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 :
  - $t \simeq 0.1066$
- Estimation au maximum de vraisemblance :
  - $\max[\ln(L(t)] \simeq -289.95$ soit $t \simeq 0.1066$



Modèles à plus d'un paramètre

- Utilisation d'une loi multinomiale $\mathcal{M}(\ell, f_{ij})$:

$$L(t) = \binom{\ell}{n_{ij}} \prod_{i,j} [f_{ij}(t)]^{n_{ij}}$$

avec n_{ij} le nombre de sites (i, j) et $f_{ij}(t)$ la probabilité d'avoir au temps t un site (i, j) dans l'alignement.

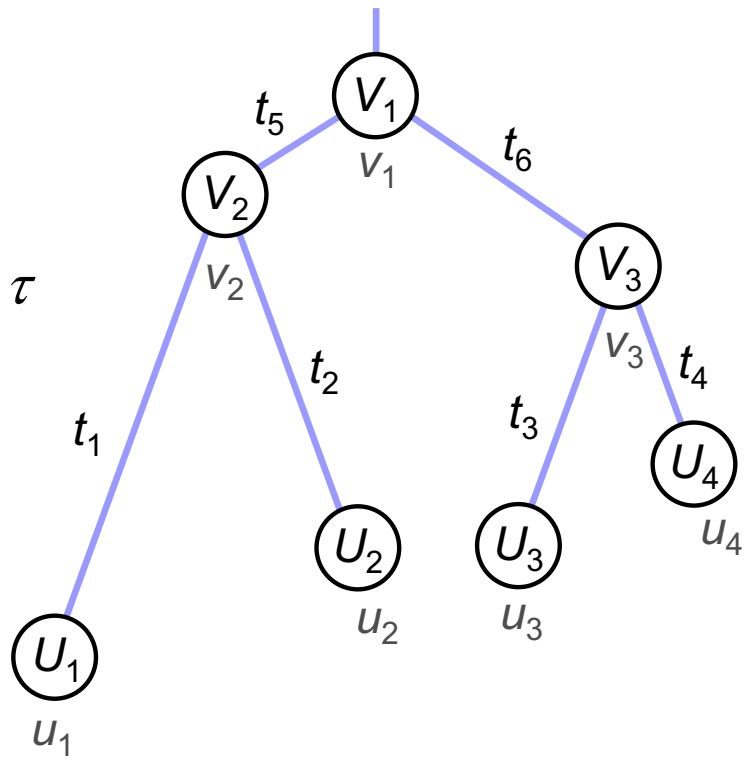
- Sous l'hypothèse de réversibilité du processus markovien, on a :

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

Simplifications possibles du fait que plusieurs valeurs de $f_{ij}(t)$ peuvent être associées à une même probabilité de transition.

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 :

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

- 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 :

$$L^{(i)}(\boldsymbol{\theta}) = \sum_{v_1}\sum_{v_2}\sum_{v_3} \mathbb{P}(v_1)\mathbb{P}(v_2|v_1, t_5)\mathbb{P}(v_3|v_1, t_6)\mathbb{P}(u_1|v_2, t_1)$$
$$\times \mathbb{P}(u_2|v_2, t_2)\mathbb{P}(u_3|v_3, t_3)\mathbb{P}(u_4|v_3, t_4)$$

Calcul de la vraisemblance

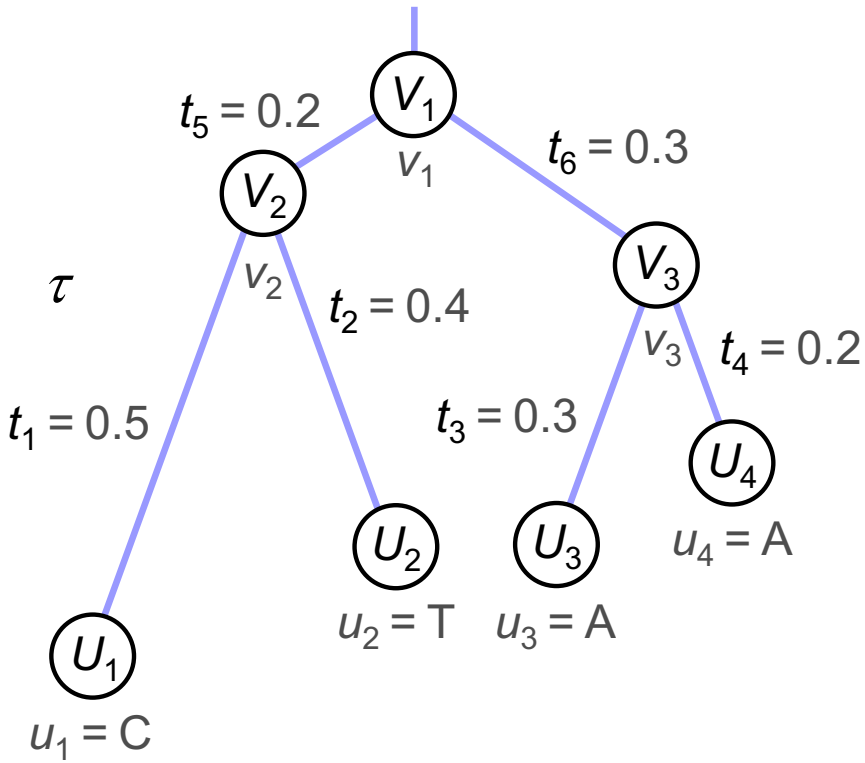
- La détermination de la vraisemblance totale nécessite le calcul de $L^{(i)}(\boldsymbol{\theta})$ pour chacun des ℓ sites.
- Le calcul des probabilités conditionnelles $\mathbb{P}(x|y,t)$ se fait par l'intermédiaire des modèles probabilistes vus précédemment.
- Sous l'hypothèse que le processus markovien 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 $\mathbf{t}$ des longueurs de branches tel que : $t_1 = 0.5$, $t_2 = 0.4$, $t_3 = t_6 = 0.3$, $t_4 = t_5 = 0.2$
  - Modèle de Jukes et Cantor à un paramètre :
    - Fréquences à l'équilibre $\pi_i = 1/4 \ \forall i$.



Probabilités de substitution

- Calcul des probabilités de substitution $i \to j$ associées à une branche de longueur t au moyen de la relation :

$$p_{ij}(t) = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4t/3} \quad (i \neq j)$$

- De la même façon, les probabilités de conservation sont :

$$p_{ii}(t) = 1 - 3p_{ij}(t) = \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4t/3}$$

- Exemples :
  - $p_{\mathrm{AT}}(t_1) = p_{\mathrm{AT}}(0.5) = 0.12$
  - $p_{\mathrm{TT}}(t_2) = p_{\mathrm{TT}}(0.4) = 0.69$
  - $p_{\mathrm{GT}}(t_3) = p_{\mathrm{GT}}(t_6) = p_{\mathrm{GT}}(0.3) = 0.08$
  - $p_{\mathrm{GA}}(t_4) = p_{\mathrm{GA}}(t_5) = p_{\mathrm{GA}}(0.2) = 0.06$

Matrices de substitution

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

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

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

Calcul d'une valeur

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

$$\begin{aligned}
&\mathbb{P}(v_1 = \mathrm{A})\mathbb{P}(v_2 = \mathrm{A}|v_1 = \mathrm{A}, t_5 = 0.2)\mathbb{P}(v_3 = \mathrm{A}|v_1 = \mathrm{A}, t_6 = 0.3) \\
&\times \mathbb{P}(u_1 = \mathrm{C}|v_2 = \mathrm{A}, t_1 = 0.5)\mathbb{P}(u_2 = \mathrm{T}|v_2 = \mathrm{A}, t_2 = 0.4) \\
&\times \mathbb{P}(u_3 = \mathrm{A}|v_3 = \mathrm{A}, t_3 = 0.3)\mathbb{P}(u_4 = \mathrm{A}|v_3 = \mathrm{A}, t_4 = 0.2)
\end{aligned}$$

Soit, avec une écriture simplifiée :

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

Calcul de toutes les combinaisons (I)

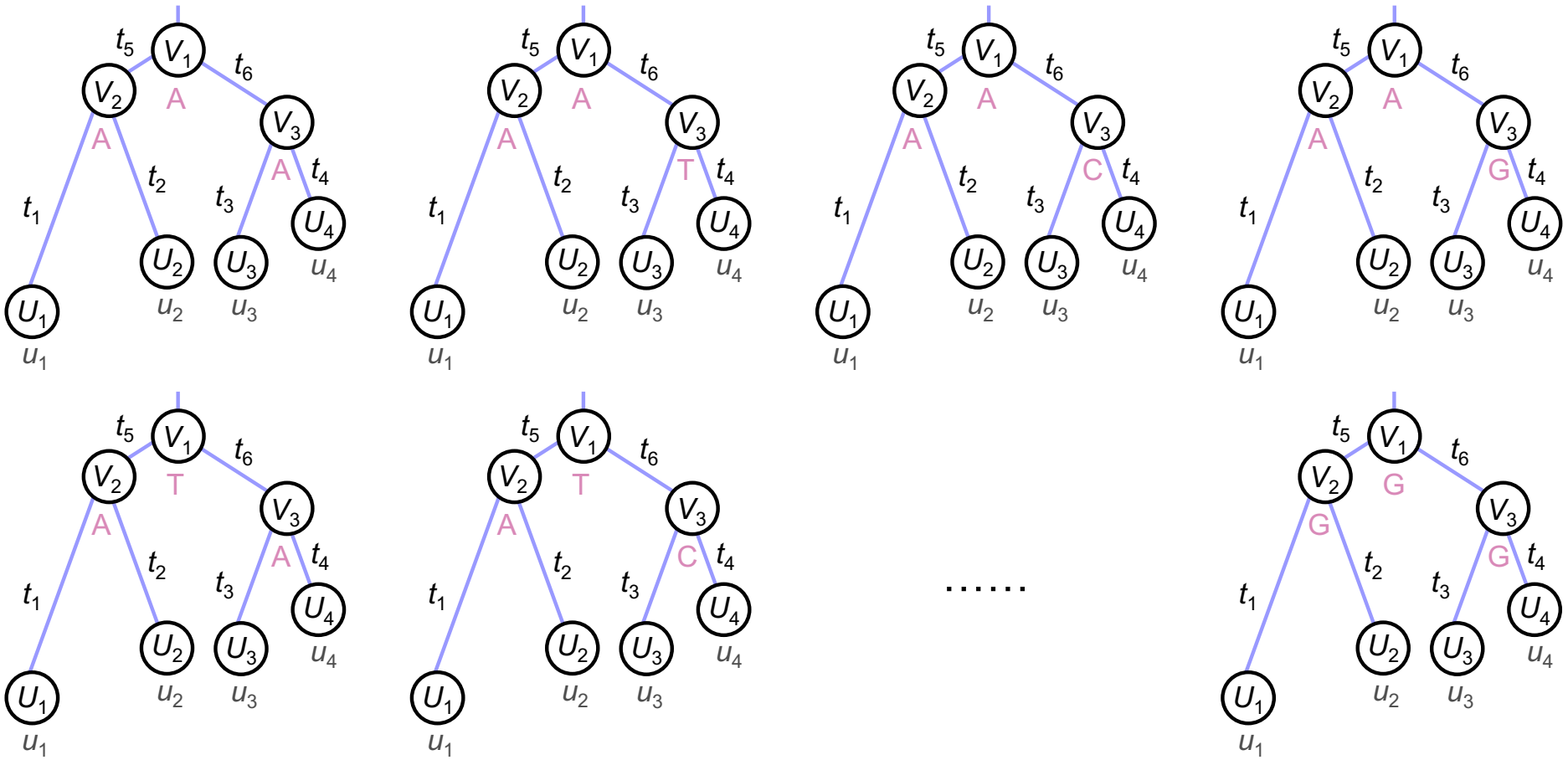
v_1	v_2	v_3	Vraisemblance	v_1	v_2	v_3	Vraisemblance
A	A	A	1.134675×10^{-3}	C	A	A	8.856×10^{-6}
A	A	C	9.4464×10^{-7}	C	A	C	6.48×10^{-7}
A	A	T	9.4464×10^{-7}	C	A	T	6.912×10^{-8}
A	A	G	9.4464×10^{-7}	C	A	G	6.912×10^{-8}
A	C	A	4.428×10^{-4}	C	C	A	6.45504×10^{-4}
A	C	C	3.6864×10^{-7}	C	C	C	4.7232×10^{-5}
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 de 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 : $L^{(i)}(\boldsymbol{\theta}) = 0.004267$

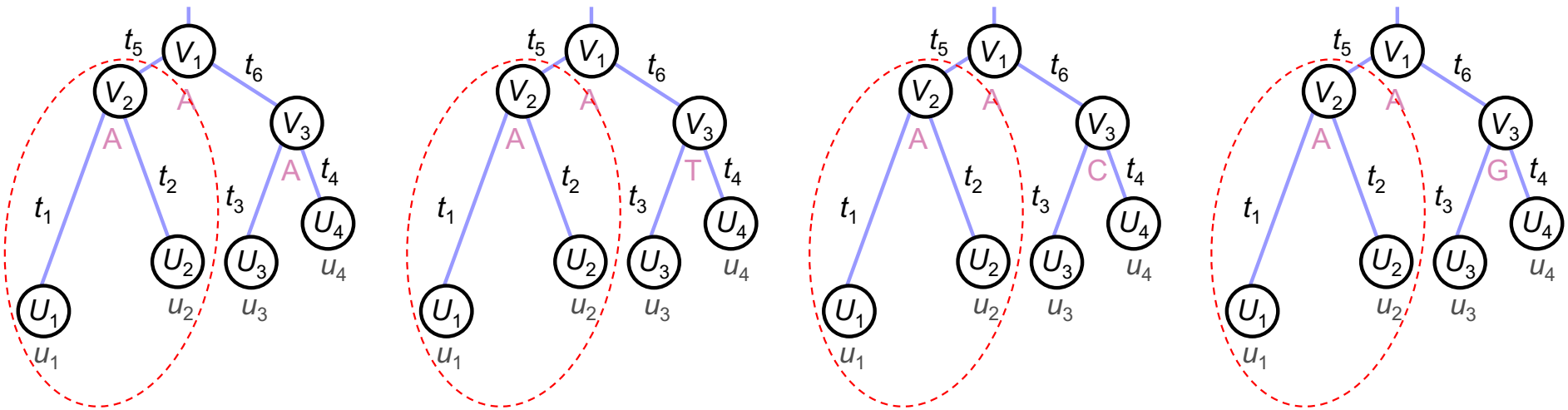
Ensemble des scénarios possibles



$4^3 = 64$ 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 d'élagage

- 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 :

$$L^{(i)}(\boldsymbol{\theta}) = \sum_{v_1} \mathbb{P}(v_1) \left[\sum_{v_2} \mathbb{P}(v_2|v_1, t_5)\mathbb{P}(u_1|v_2, t_1)\mathbb{P}(u_2|v_2, t_2) \right] \times \left[\sum_{v_3} \mathbb{P}(v_3|v_1, t_6)\mathbb{P}(u_3|v_3, t_3)\mathbb{P}(u_4|v_3, t_4) \right]$$

  - 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'arbre, 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 ambiguïtés 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

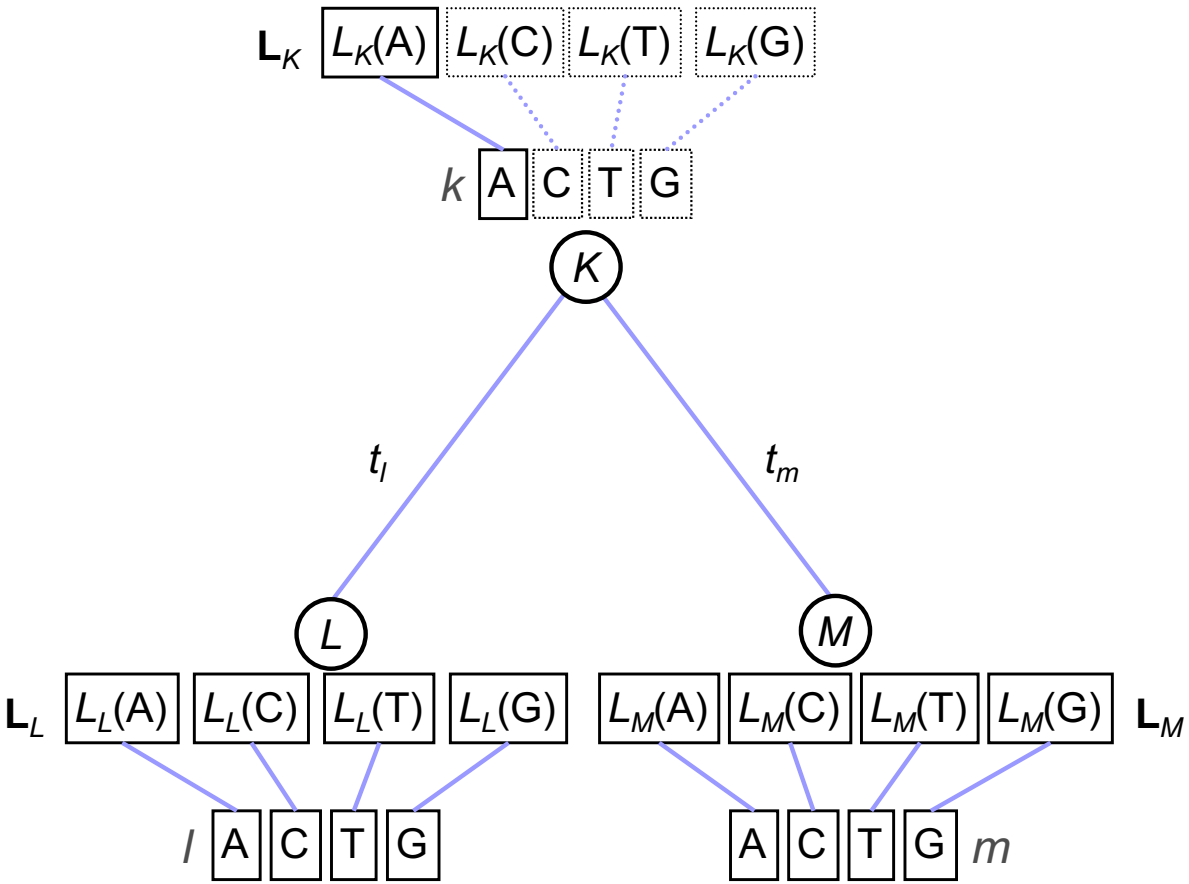
- Si K correspond à un nœud, alors :

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

  avec L et M les deux nœuds fils de K, t_l la longueur de la branche reliant K à L et t_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.

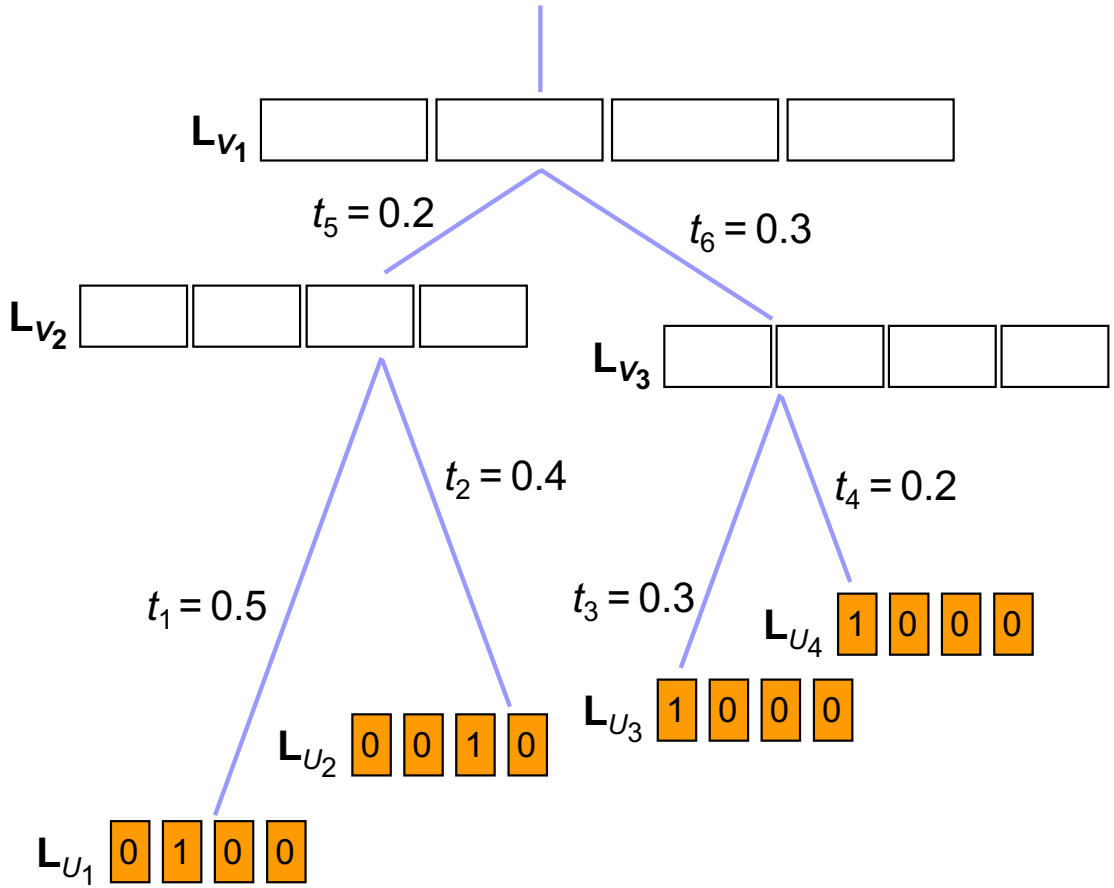
Vraisemblances partielles aux feuilles

- 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 $L_{V_2}^{(i)}(\mathrm{T})$:

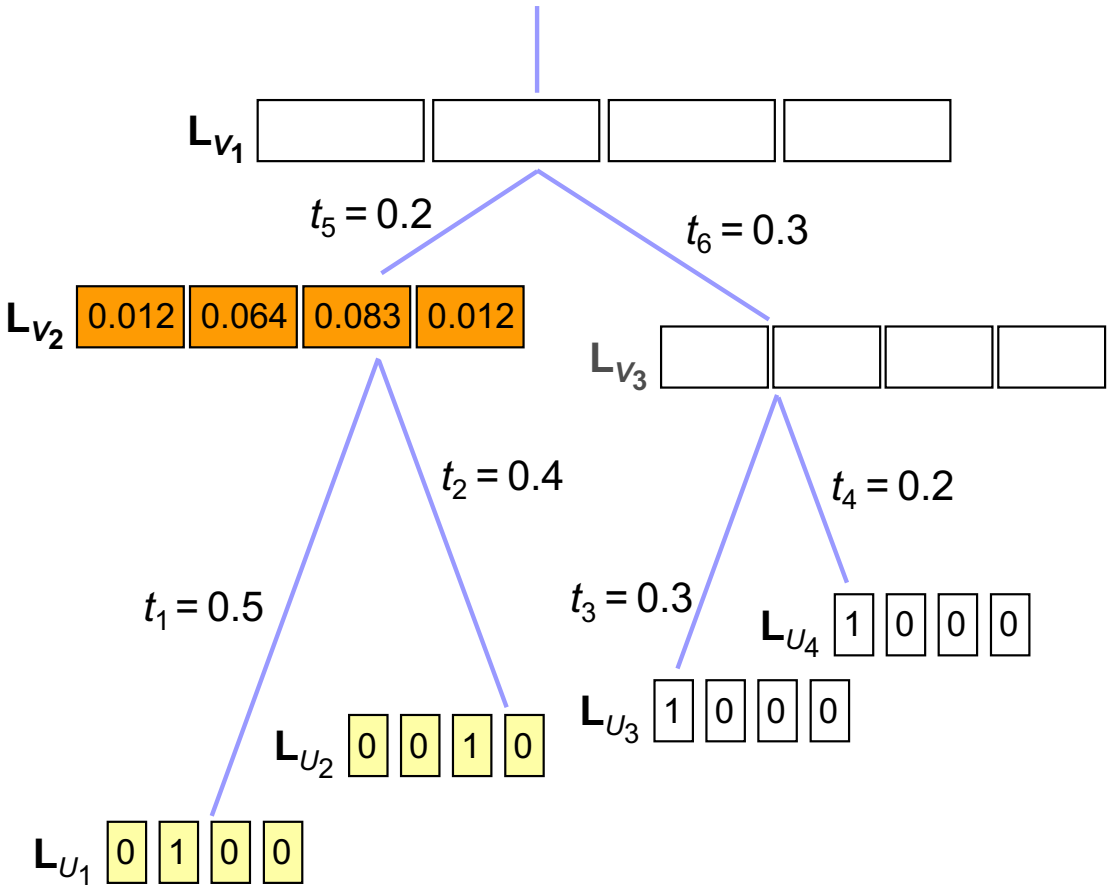
$$\begin{aligned}
L_{V_2}^{(i)}(\mathrm{T}) &= \left[p_{\mathrm{TA}}(0.5)L_{U_1}^{(i)}(\mathrm{A}) + p_{\mathrm{TC}}(0.5)L_{U_1}^{(i)}(\mathrm{C}) + p_{\mathrm{TT}}(0.5)L_{U_1}^{(i)}(\mathrm{T}) + p_{\mathrm{TG}}(0.5)L_{U_1}^{(i)}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{TA}}(0.4)L_{U_2}^{(i)}(\mathrm{A})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

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

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

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

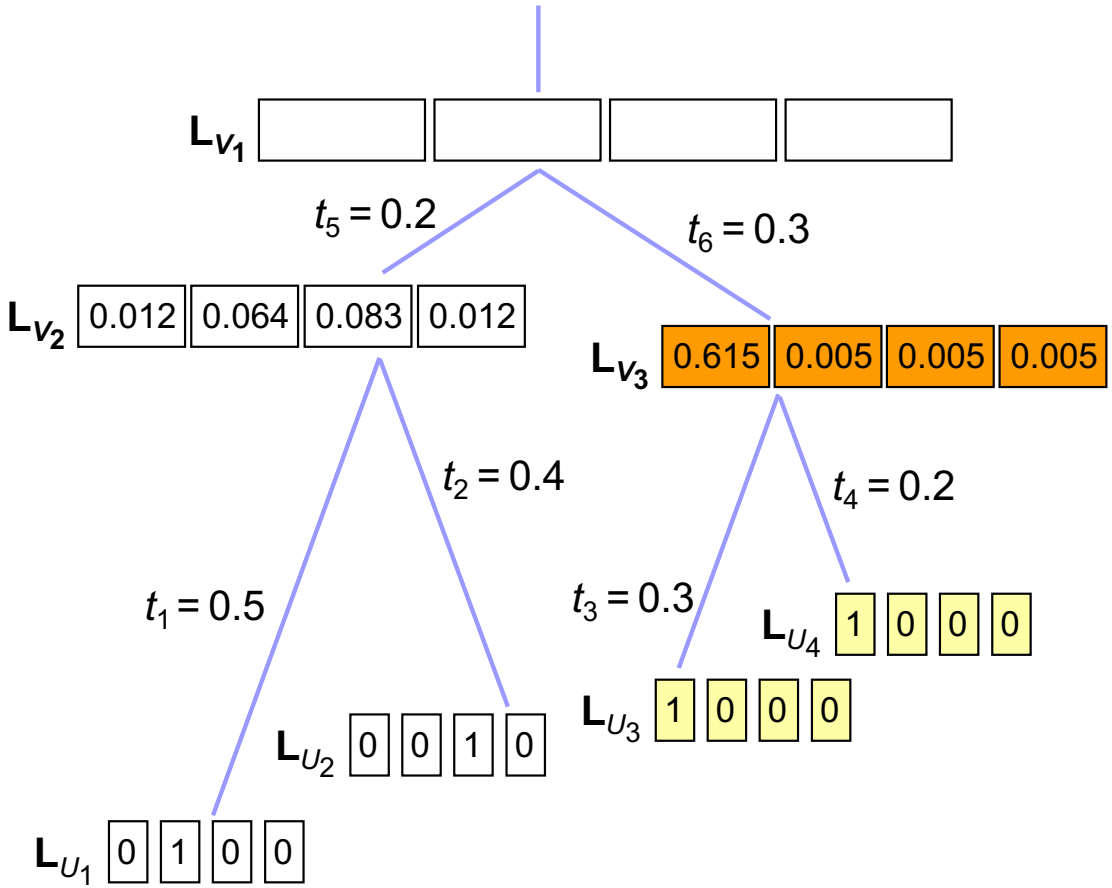
$$\begin{aligned}
L^{(i)}_{V_3}(\mathrm{T}) &= \Big[p_{\mathrm{TA}}(0.3)L^{(i)}_{U_3}(\mathrm{A}) + p_{\mathrm{TC}}(0.3)L^{(i)}_{U_3}(\mathrm{C}) + p_{\mathrm{TT}}(0.3)L^{(i)}_{U_3}(\mathrm{T}) + p_{\mathrm{TG}}(0.3)L^{(i)}_{U_3}(\mathrm{G})\Big] \\
&\times \Big[p_{\mathrm{TA}}(0.2)L^{(i)}_{U_4}(\mathrm{A})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

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

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

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

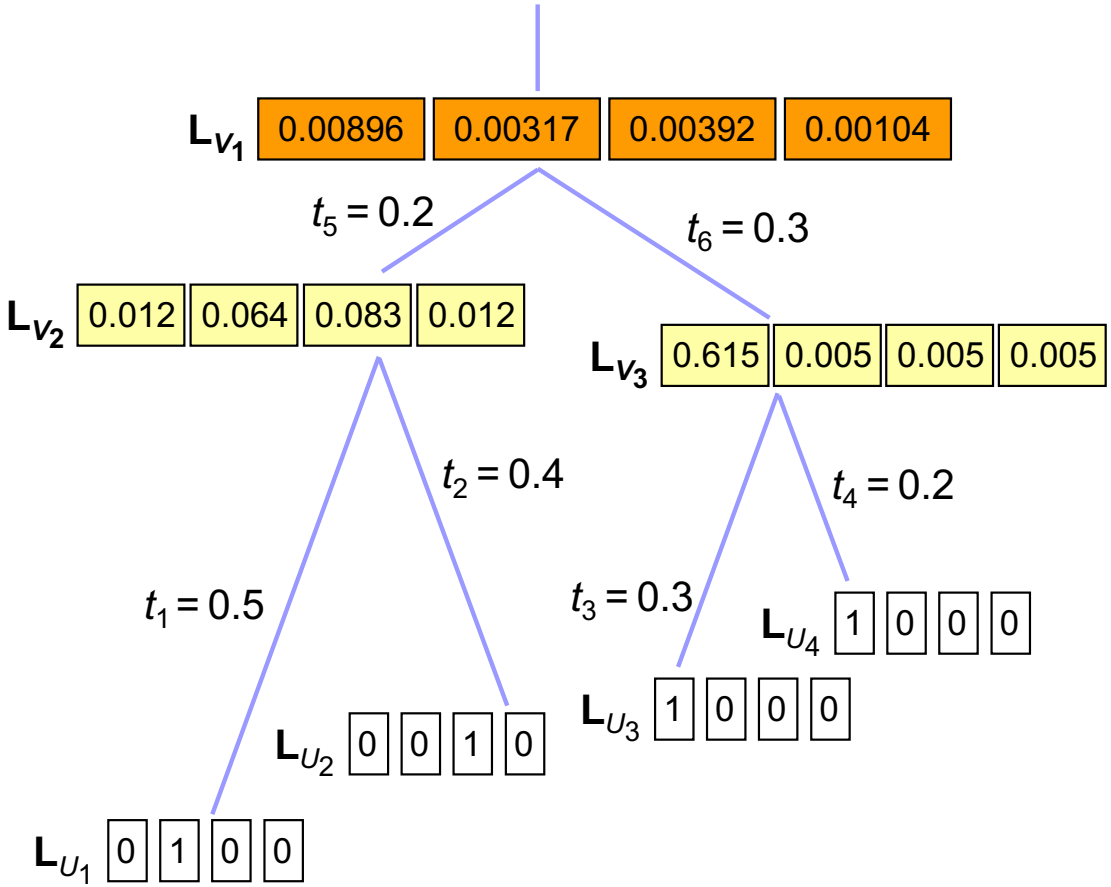
$$
\begin{aligned}
L_{V_1}^{(i)}(\mathrm{T}) &= \left[p_{\mathrm{TA}}(0.2)L_{V_2}^{(i)}(\mathrm{A}) + p_{\mathrm{TC}}(0.2)L_{V_2}^{(i)}(\mathrm{C}) + p_{\mathrm{TT}}(0.2)L_{V_2}^{(i)}(\mathrm{T}) + p_{\mathrm{TG}}(0.2)L_{V_2}^{(i)}(\mathrm{G})\right] \\
&\times \left[p_{\mathrm{TA}}(0.3)L_{V_3}^{(i)}(\mathrm{A})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

- 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 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

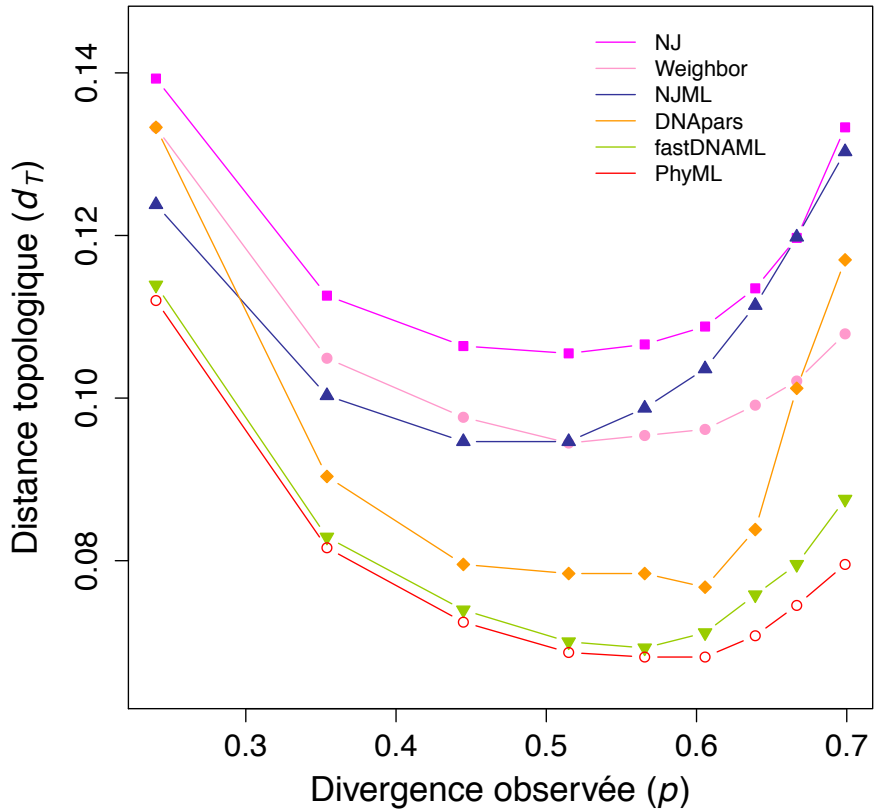
- En théorie, nécessité d'explorer l'ensemble des topologies et des combinaisons de longueurs de branches :
  - Impossible du fait de la croissance très rapide du nombre de topologies et du caractère continu des longueurs de branches.
- En pratique :
  - Exploration de l'espace des topologies via les heuristiques vues précédemment.
  - 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 (surtout si *bootstrap*).
- Impossibilité d'explorer l'ensemble des topologies lorsque $n \geq 12$:
  - L'emploi d'heuristiques fait que l'on est pas sûr d'avoir l'arbre le plus vraisemblable.
- Risques de surparamétrisation avec les modèles trop complexes :
  - Tests pour sélectionner le modèle permettant d'obtenir le meilleur compromis vraisemblance/nb. de paramètres (LRT, AIC, BIC).

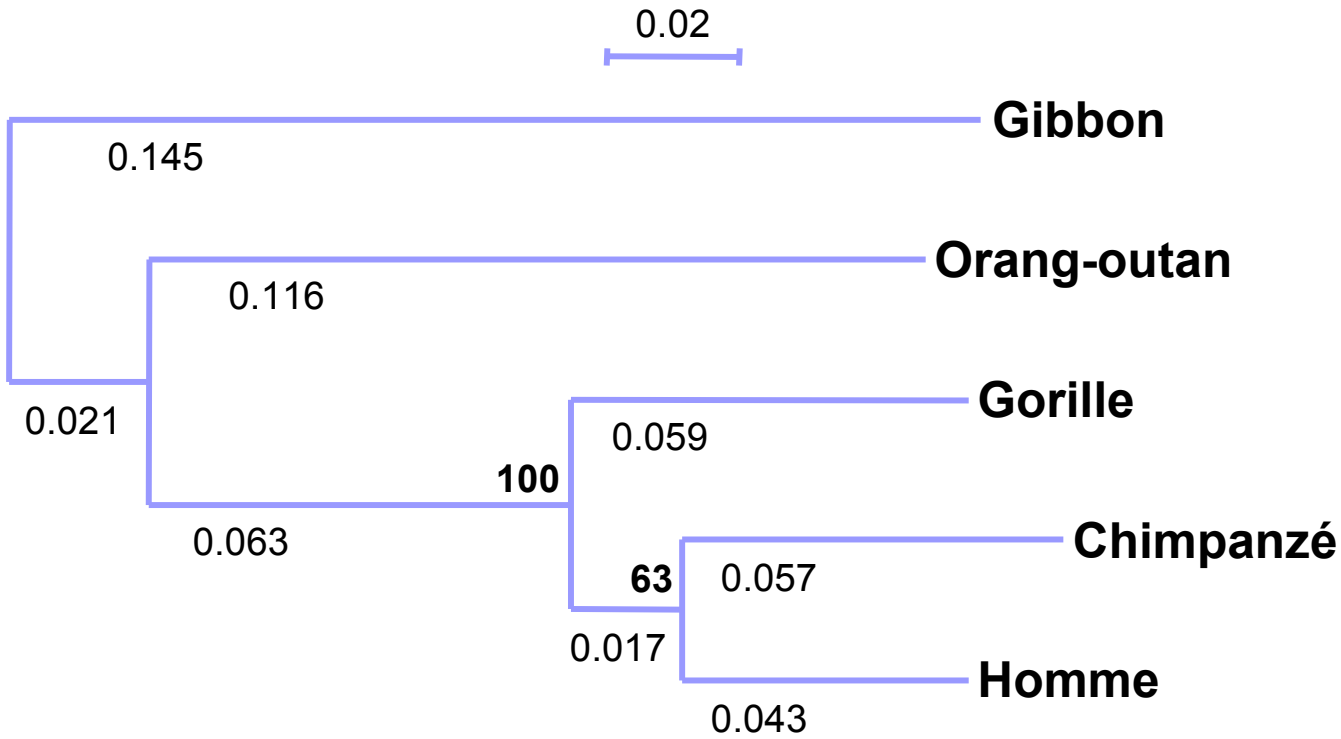
Performances en simulation

- 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.

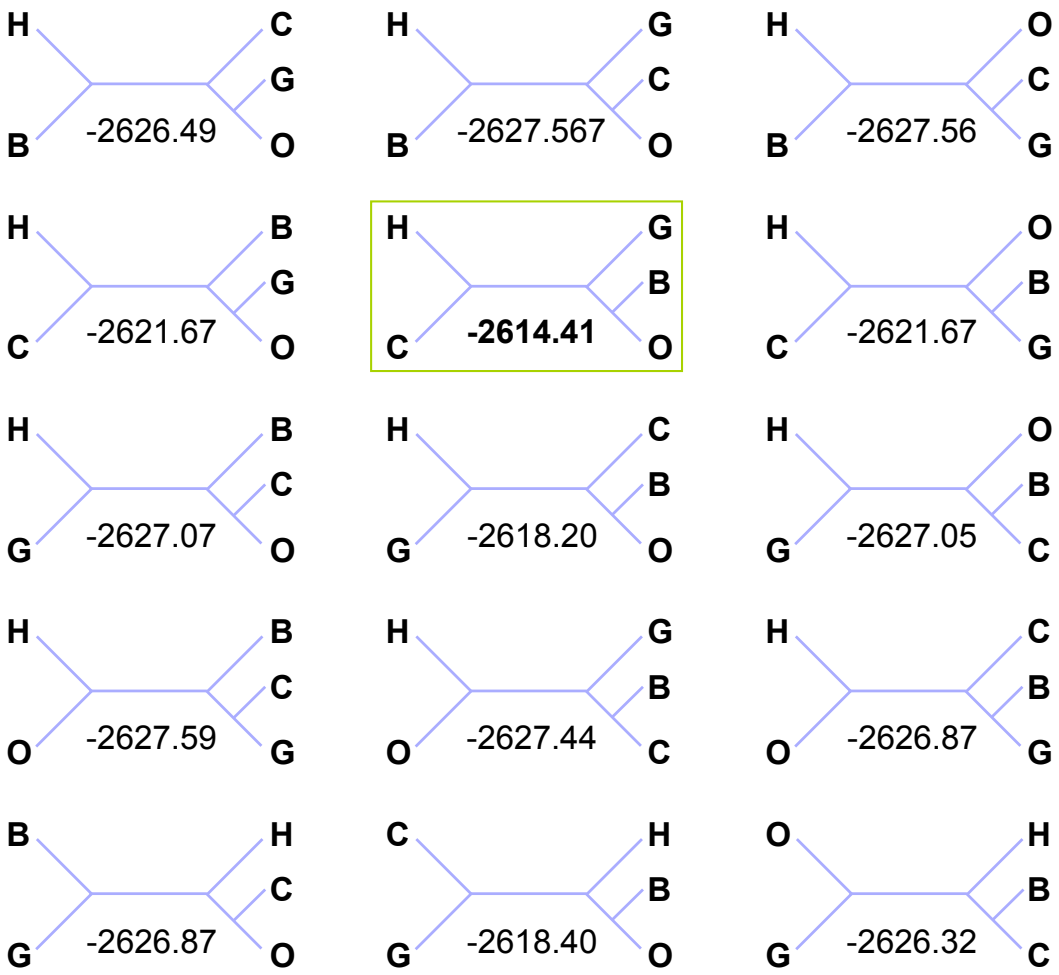


Phylogénie des Hominoïdes

- Sélection du modèle HKY+Γ après un test BIC.
- Racinement avec la séquence du Gibbon.
- 500 réplicats de *bootstrap*.



Vraisemblances des topologies

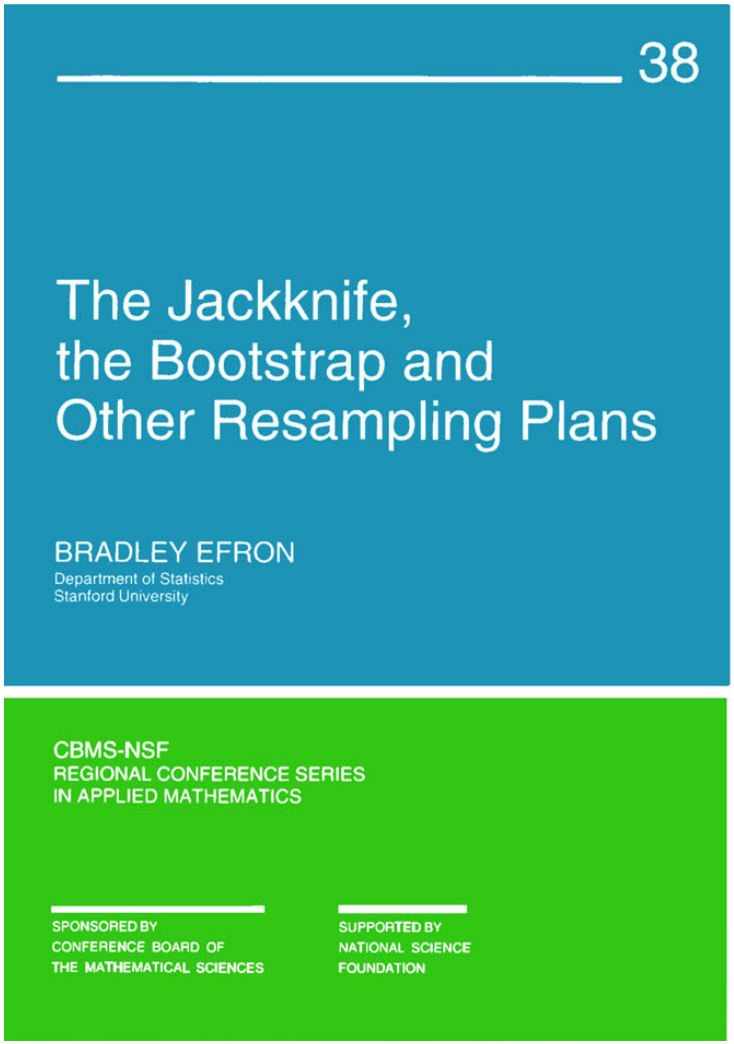


B = Gibbon, H = Homme, C = Chimpanzé, G = Gorille, O = Orang-outan

Plan

Le *bootstrap*

- Bases mathématiques établies par Efron (1979) :
  - Construction d'intervalles de confiance.
  - Mesure de la précision d'une estimation.
- Adaptation à la phylogénie par Felsenstein (1985) :
  - Méthode aujourd'hui la plus couramment utilisée.

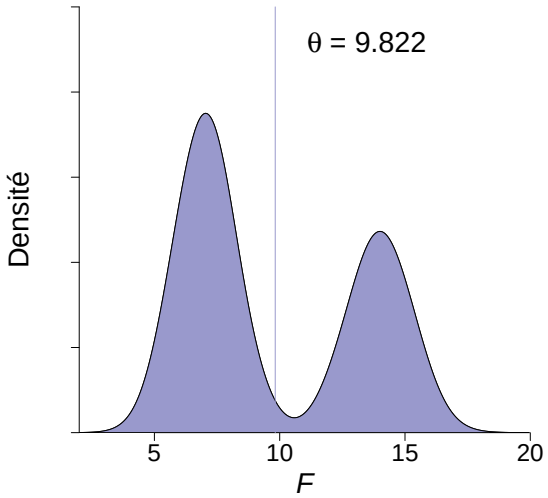


Principe général

- Soit un échantillon $\mathbf{x} = (x_1, x_2, \ldots, x_\ell)$ de ℓ observations tirées d'une distribution $\mathcal{F}$, de paramètre θ inconnu :
  - Soit $\hat{\mathcal{F}}$ la distribution observée dans cet échantillon :
    - Estimation de θ à partir de $\hat{\mathcal{F}}$.
- Mesure de l'intervalle de confiance de l'estimation précédente au moyen du *bootstrap* :
  - Tirage de B échantillons $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_\ell^*)$ à partir de $\hat{\mathcal{F}}$.
  - Chaque $\mathbf{x}^*$ est construit par ℓ tirages avec remise dans $\mathbf{x}$ et constitue ce que l'on appelle un *réplicat de bootstrap*.
  - $I(\theta)$ à 95% obtenu en retirant les 2.5% de valeurs les plus hautes et les 2.5% de valeurs les plus basses.
  - Nécessité que B et ℓ soient grands et que les observations de $\mathbf{x}$ soient i.i.d.

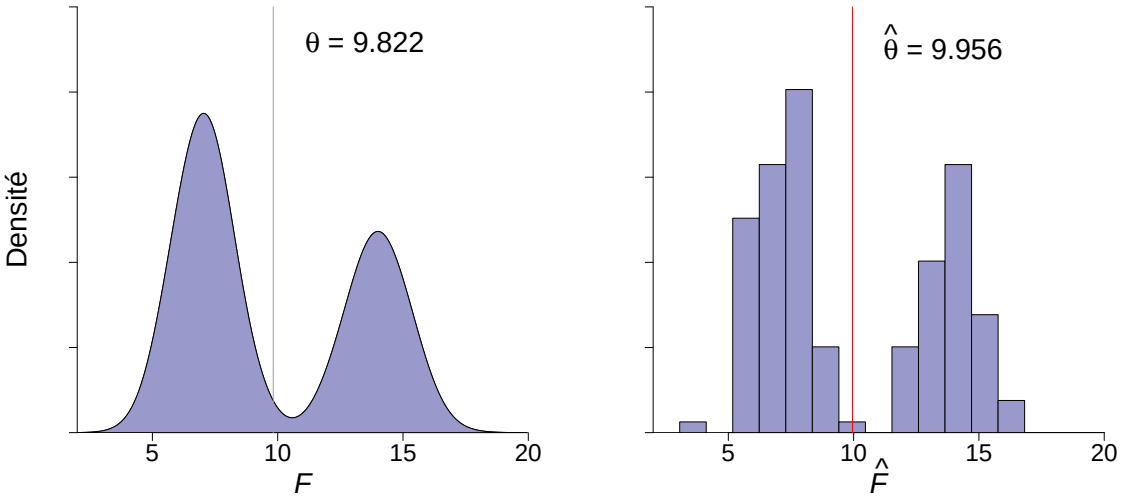
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7, 1)$ pour 60% des effectifs et $\mathcal{N}(14, 1)$ pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.



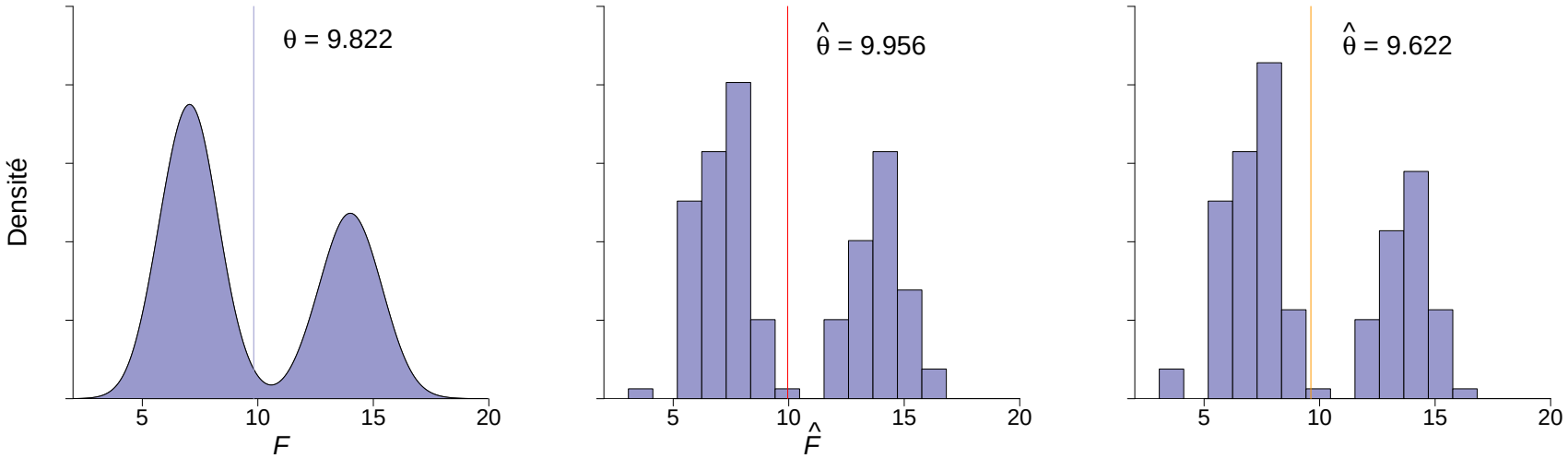
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7, 1)$ pour 60% des effectifs et $\mathcal{N}(14, 1)$ pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.



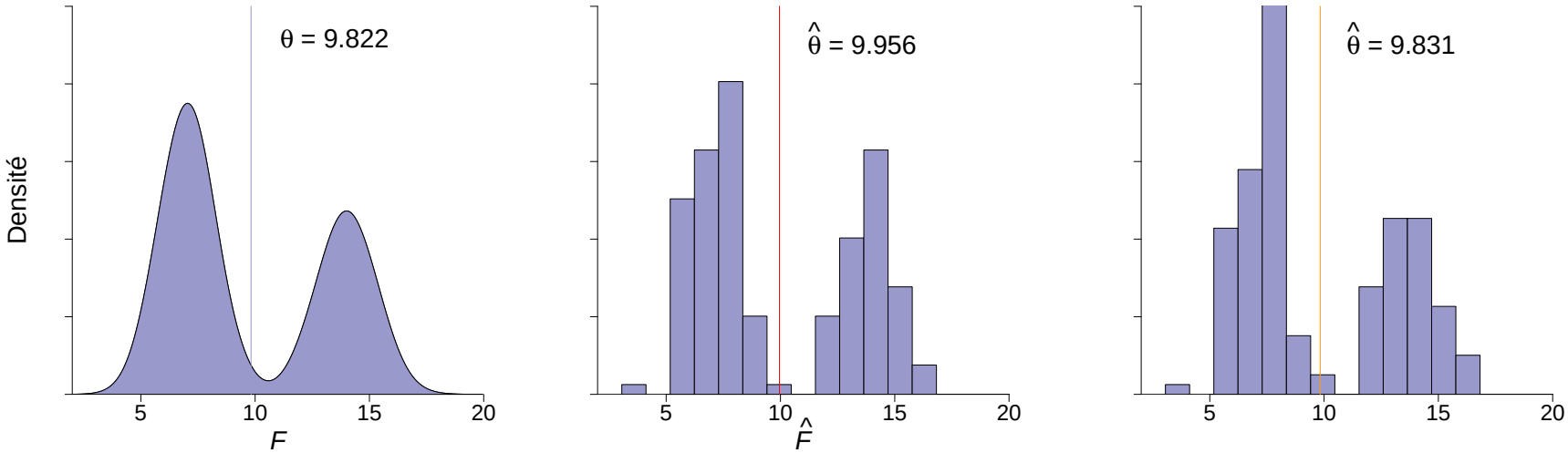
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7, 1)$ pour 60% des effectifs et $\mathcal{N}(14, 1)$ pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :



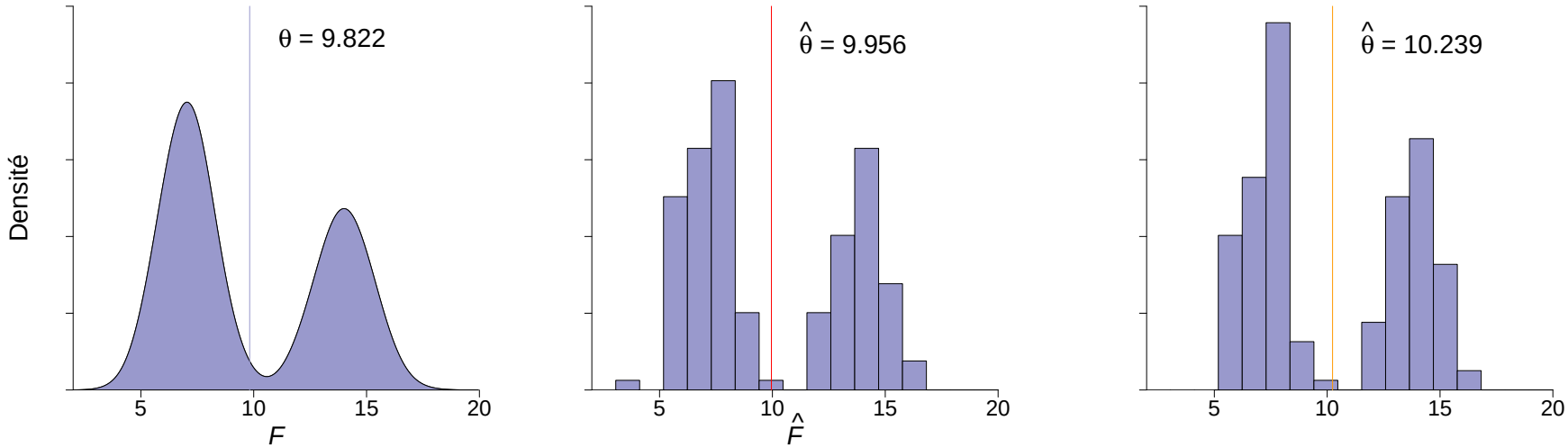
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7, 1)$ pour 60% des effectifs et $\mathcal{N}(14, 1)$ pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :



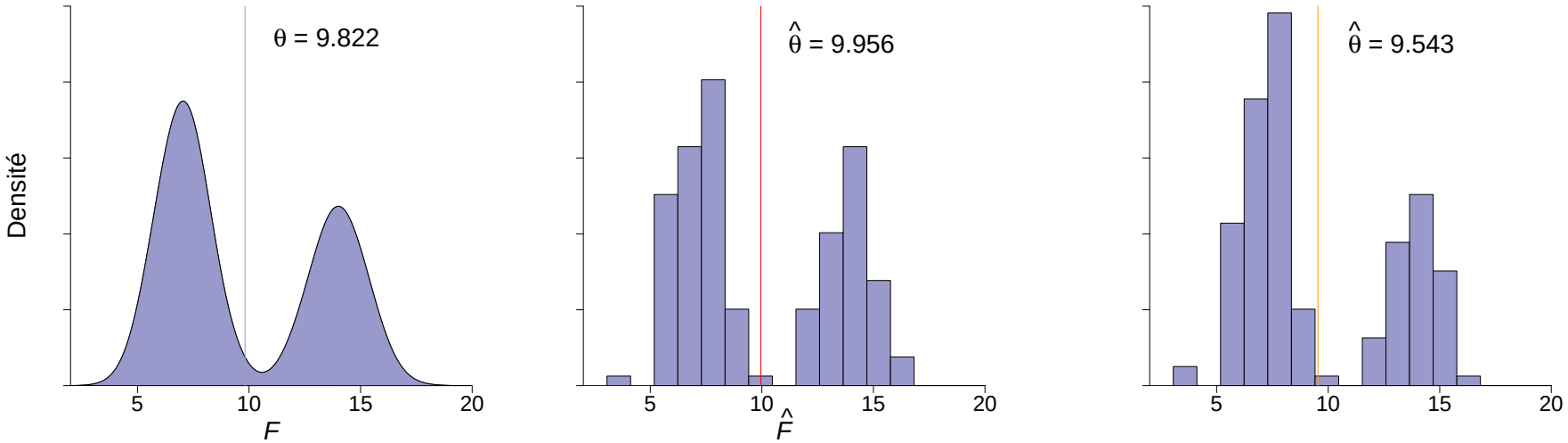
Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7, 1)$ pour 60% des effectifs et $\mathcal{N}(14, 1)$ pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :

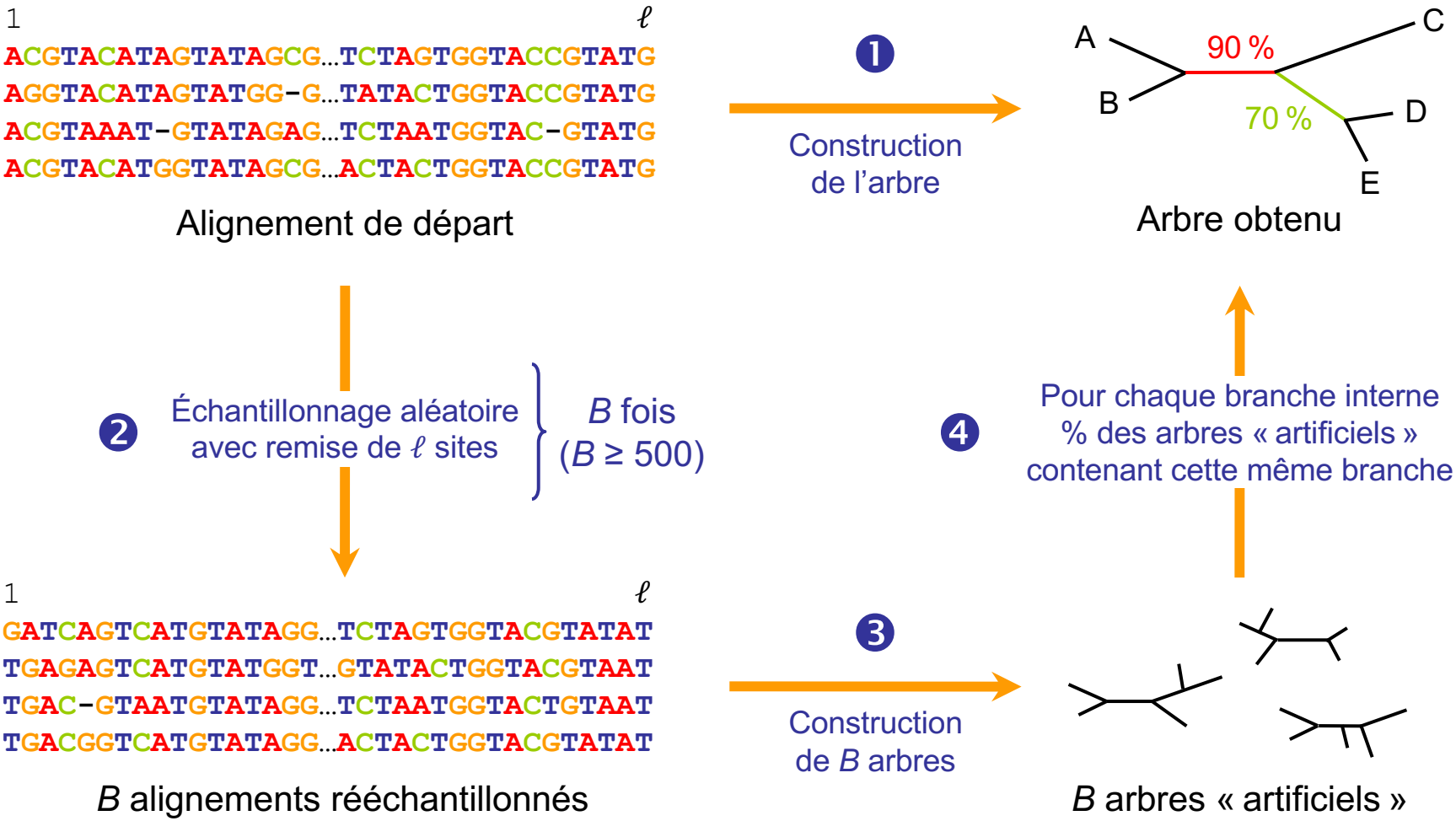


Moyenne d'une distribution

- Construction d'une distribution $\mathcal{F}$ par le mélange de deux lois normales :
  - $\mathcal{N}(7, 1)$ pour 60% des effectifs et $\mathcal{N}(14, 1)$ pour 40% des effectifs :
    - Moyenne de la distribution : $\theta = 9.822$.
- Tirage de $\ell = 150$ individus dans $\mathcal{F}$ pour construire $\hat{\mathcal{F}}$:
  - Moyenne estimée : $\hat{\theta} = 9.956$.
  - Mesure de la validité de cette estimation par *bootstrap* :



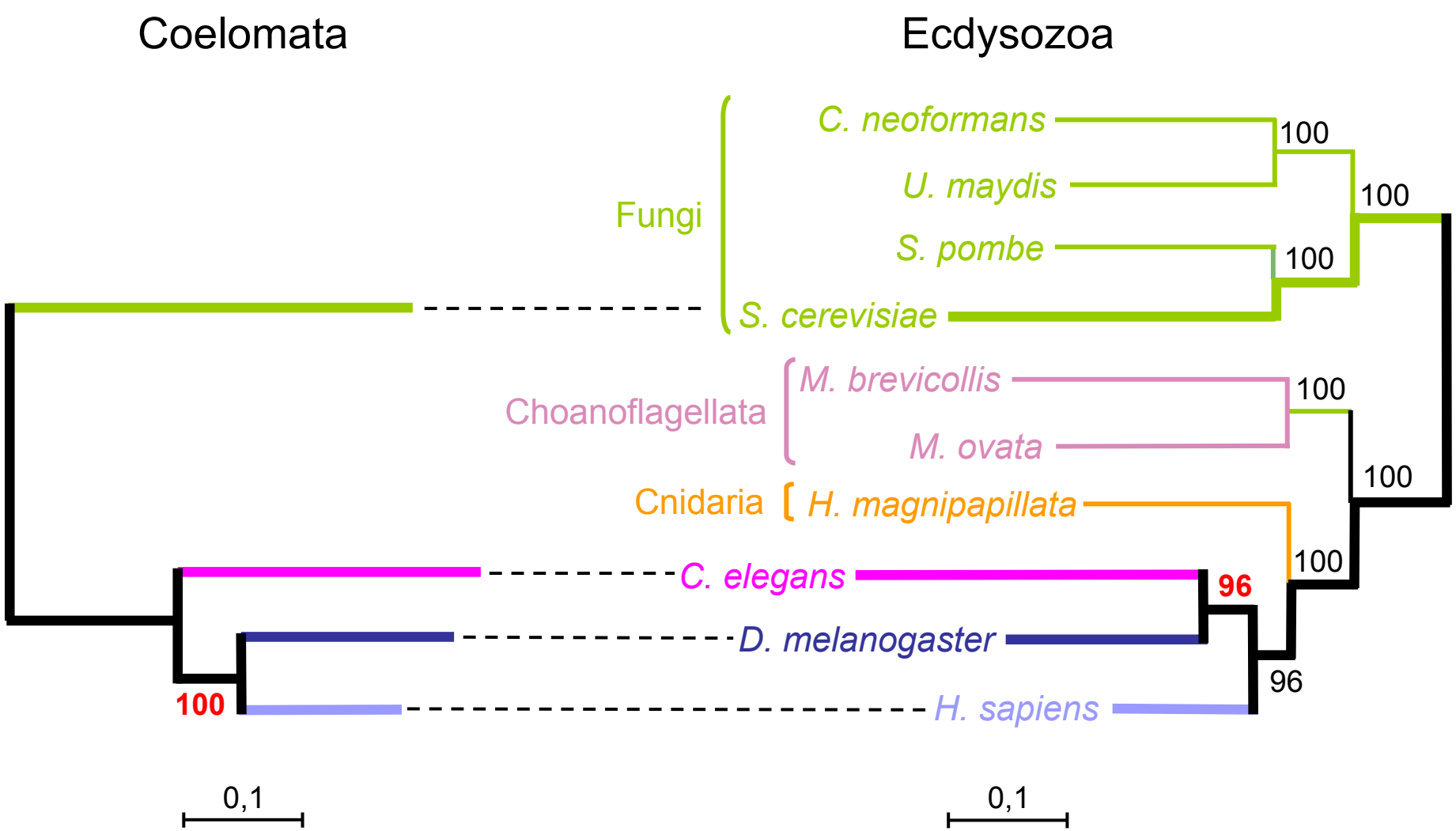
Application à la phylogénie



Limitations et usage

- Ne permet pas de déterminer si un arbre est vrai ou faux :
  - Un arbre faux peut avoir des branches soutenues par de fortes valeurs de *bootstrap*.
- Non-indépendance des observations (sites) :
  - Surestimation des scores faibles et sous-estimation des scores forts.
- En théorie, seuil en fonction d'un risque d'erreur fixé *a priori* :
  - En pratique, valeurs fluctuantes suivant les utilisateurs.
  - Seuils communément admis :
    - 100% : robustesse maximale.
    - 95-99% : très fort soutien par les données.
    - 90-94% : fort soutien par les données.
    - 80-89% : soutien modéré par les données.
    - $< 80\%$: pas de soutien.

Un exemple classique



Approximate Likelihood Ratio Test (aLRT)

- Alternative à l'utilisation du *bootstrap*, très coûteux en temps de calcul dans le cas du maximum de vraisemblance.
- Calcul de la statistique :
  - Soit τ_1 la topologie présentant la vraisemblance maximale $L(\tau_1)$.
  - Soit τ_2 la topologie présentant la *deuxième* vraisemblance maximale $L(\tau_2)$:
    - Obtention par réarrangement NNI autour de la branche d'intérêt b_k.
    - Fixation des autres paramètres $(\mathbf{t}, \boldsymbol{\vartheta}, \alpha)$.
  - Le rapport des vraisemblances est donné par :

$$\Lambda_k = 2 \ln \left[\frac{L(\tau_1)}{L(\tau_2)} \right] = 2 \left[\ln L(\tau_1) - \ln L(\tau_2) \right]$$

- Calcul du test :

$$\Lambda_k \sim \frac{1}{2} \left[\chi^2(0) + \chi^2(1) \right]$$

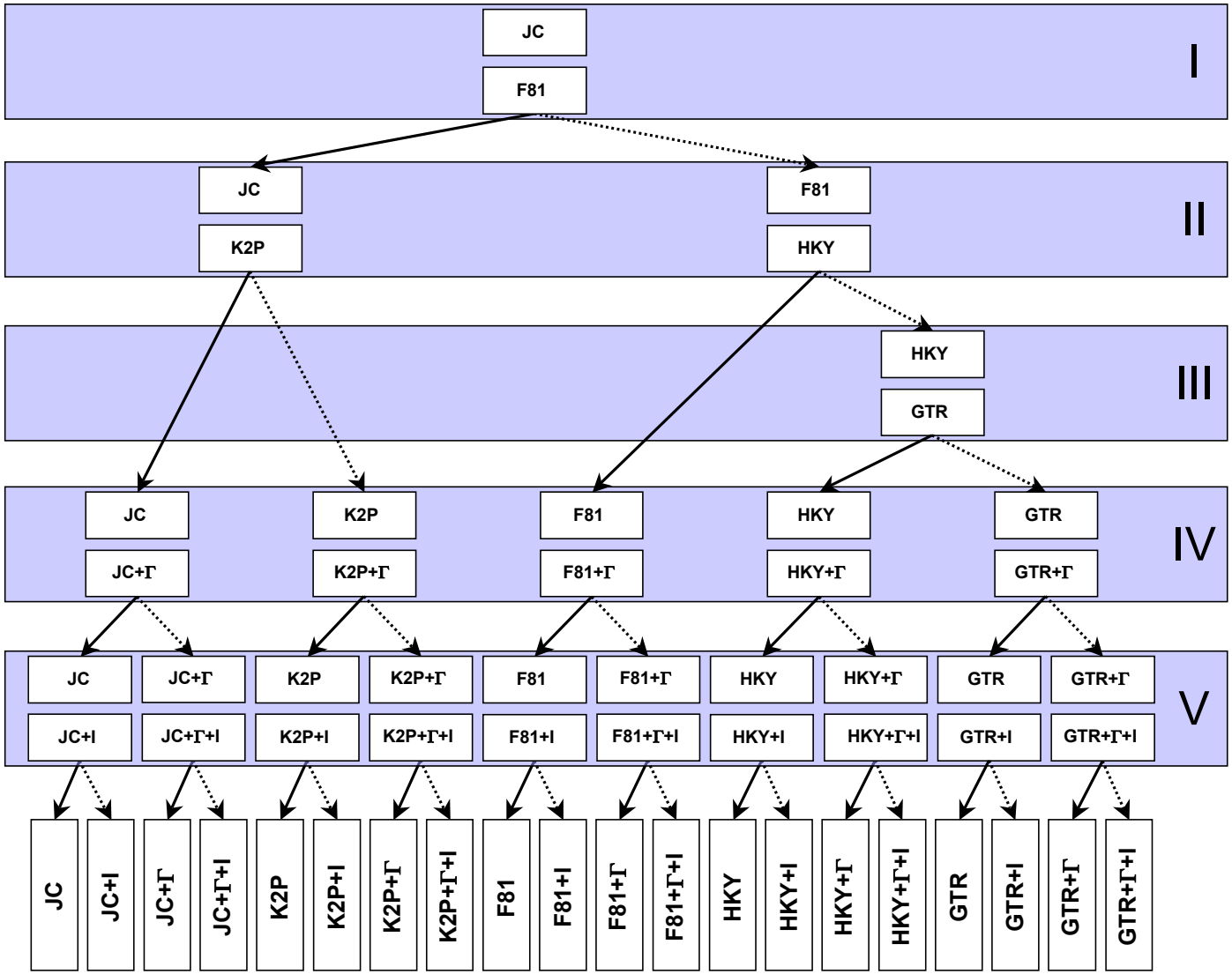
Likelihood Ratio Test (LRT)

- 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 :

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{\ell - k - 1}$$

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 :

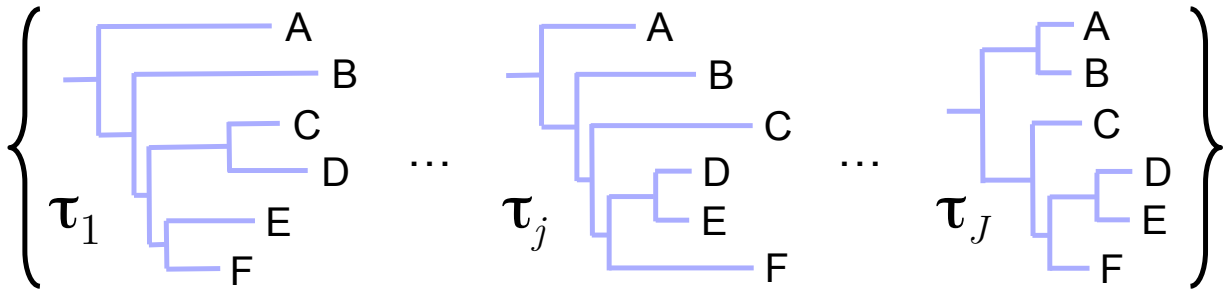
$$\mathrm{BIC} = -2 \ln L(\boldsymbol{\vartheta}) + k \ln \ell$$

- 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 (*cf.* cours sur l'inférence bayésienne) :

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

Nécessité d'utilisation

- Différents jeux de données peuvent retourner différents arbres.
- Différentes méthodes peuvent retourner différents arbres.
- Une même méthode peut retourner différents arbres.
- Les différences observées sont-elles significatives ?



Utilisation de tests de vraisemblance

Tests courants

- Kishino et Hasegawa (KH – Kishino et Hasegawa, 1989).
- Shimodaira et Hasegawa (SH – Shimodaira et Hasegawa, 1999).
- *Expected Likelihood Weight* (ELW – Strimmer et Rambaut, 2001).
- *Approximately Unbiased* (AU – Shimodaira, 2002).

Test de Kishino et Hasegawa

- Soit S un alignement de séquences de longueur ℓ et $L(\boldsymbol{\theta}_1)$ et $L(\boldsymbol{\theta}_2)$ les vraisemblances de deux arbres obtenus à partir de S.
- On pose $Y_1 = \ln L(\boldsymbol{\theta}_1)$ et $Y_2 = \ln L(\boldsymbol{\theta}_2)$ et $\Delta = Y_1 - Y_2$.
- Le test KH consiste à tester si Δ est significativement différent de zéro, ce qui revient à la formulation :

$$H_0 : \mathbb{E}(\Delta) = 0$$
$$H_1 : \mathbb{E}(\Delta) \neq 0$$

- Le problème est que la distribution de Δ n'est pas connue :
  - Estimation de la variance de Δ au moyen de différentes méthodes.

Approche classique (I)

- Soit $y_1^{(i)} = \ln L^{(i)}(\boldsymbol{\theta}_1)$ et $y_2^{(i)} = \ln L^{(i)}(\boldsymbol{\theta}_2)$, dans ce cas les valeurs de Y_1 et Y_2 sont telles que :

$$Y_1 = \sum_{i=1}^{\ell} y_1^{(i)} \quad \text{et} \quad Y_2 = \sum_{i=1}^{\ell} y_2^{(i)}$$

- Soit $\delta^{(i)} = y_1^{(i)} - y_2^{(i)}$, la différence des valeurs de vraisemblance par site, dans ce cas :

$$\Delta = Y_1 - Y_2 = \sum_{i=1}^{\ell} \delta^{(i)}$$

Approche classique (II)

- La moyenne des différences des valeurs de vraisemblances est donc égale à :

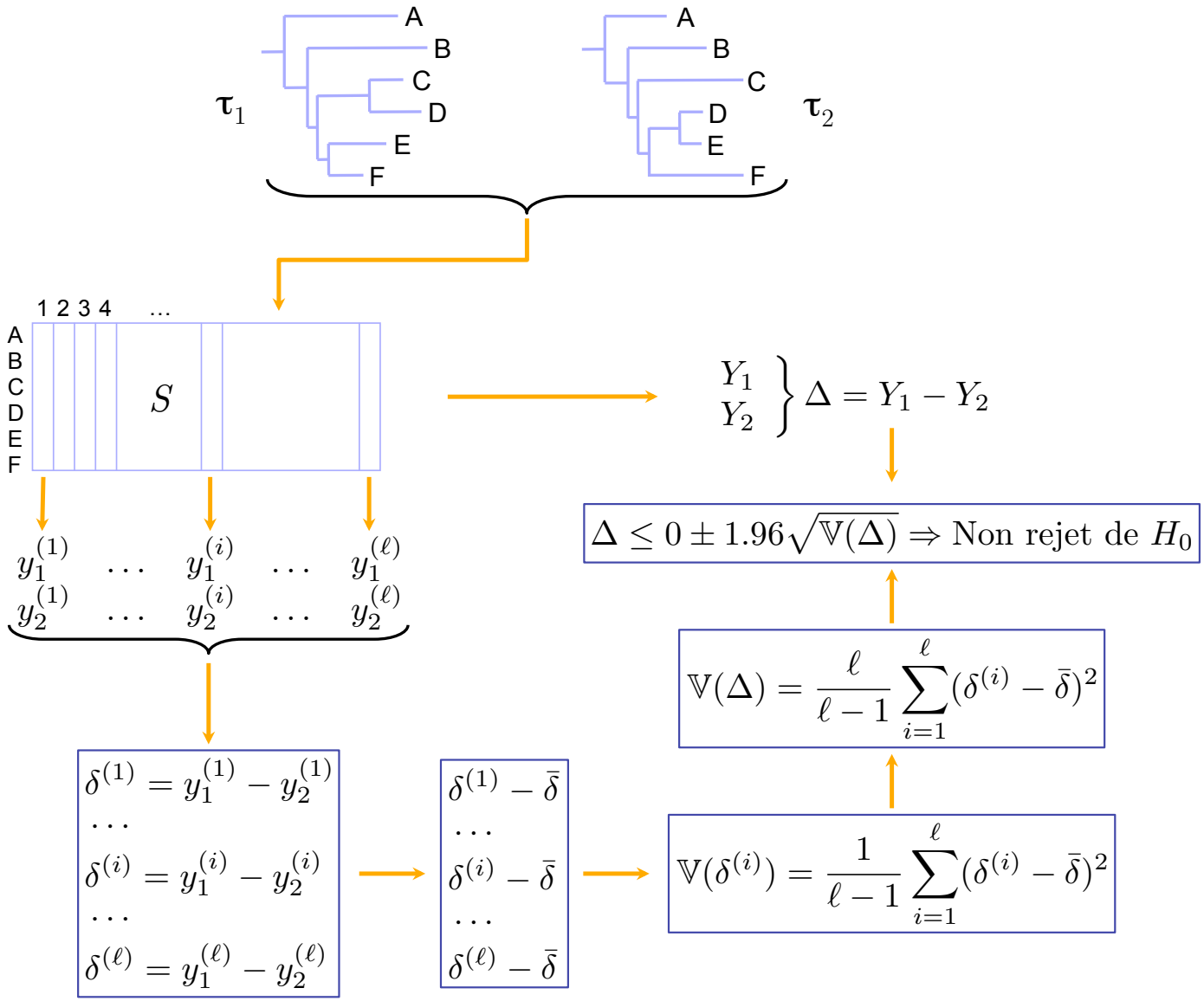
$$\bar{\delta} = \frac{1}{\ell} \sum_{i=1}^{\ell} \delta^{(i)} = \frac{\Delta}{\ell}$$

- Estimation de la variance de Δ par :

$$\mathbb{V}(\Delta) = \mathbb{V}\left(\delta^{(i)}\right) = \frac{1}{\ell - 1} \sum_{i=1}^{\ell} \left(\delta^{(i)} - \bar{\delta}\right)^2$$

- Utilisation de cette estimation pour réaliser un test bilatéral sous l'hypothèse que $\Delta \sim \mathcal{N}(0, \mathbb{V}(\Delta))$.

Schéma général



Approche par *bootstrap* (I)

- Réalisation de B rééchantillonnages des sites de S par une approche de type *bootstrap*.
- Calcul, pour chaque réplicat k ($1 \leq k \leq B$), des vraisemblances *approchées* $Y'_{1(k)}$ et $Y'_{2(k)}$ associées aux topologies τ_1 et τ_2 :
  - Utilisation des valeurs de vraisemblances par sites provenant de S pour effectuer ce calcul.
- Calcul pour chaque réplicat de $\Delta'_{(k)} = Y'_{1(k)} - Y'_{2(k)}$.
- La moyenne des valeurs de $\Delta'_{(k)}$ est telle que :

$$\bar{\Delta}' = \frac{1}{B} \sum_{k=1}^{B} \Delta'_{(k)}$$

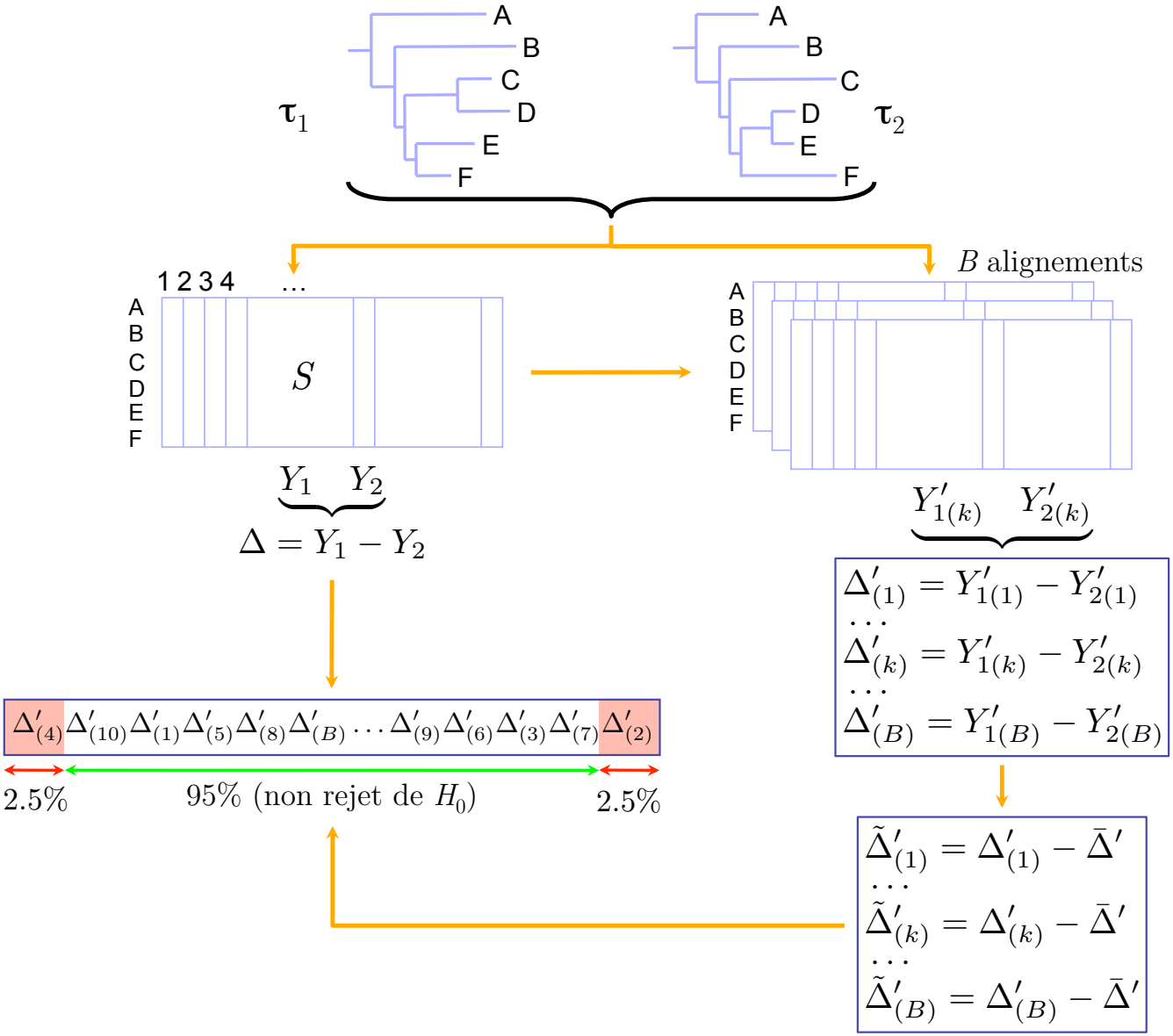
Approche par *bootstrap* (II)

- Calcul des valeurs de $\Delta'_{(k)}$ centrées par la moyenne :

$$\tilde{\Delta}'_{(k)} = \Delta'_{(k)} - \bar{\Delta}'$$

- Estimation de la variance de Δ par celle de $\tilde{\Delta}'_{(k)}$.
- Utilisation de cette variance pour réaliser un test bilatéral sous l'hypothèse que $\Delta \sim \mathcal{N}\left(0, \mathbb{V}\left(\tilde{\Delta}'_{(k)}\right)\right)$.
- Une autre possibilité est la comparaison directe de Δ avec la distribution des $\tilde{\Delta}'_{(k)}$.

Schéma général



Limitations

- Test limité à la comparaison de deux topologies :
  - Pas de correction pour les tests multiples.
- Les arbres testés doivent être choisis *indépendamment* des données utilisées pour réaliser le test :
  - Indispensable pour justifier l'hypothèse nulle sous laquelle $\mathbb{E}(\Delta) = 0$.
  - Le choix ne peut donc pas se faire sur la base de la vraisemblance.
- A malheureusement été fréquemment utilisé en violation de ces deux conditions !
- Les autres méthodes (SH, AU, ELW) utilisent un principe similaire mais corrigent ces défauts.

Phylogénie des Hominoïdes

j	τ_j	Y_j	Δ	KH	SH	ELW	AU
1	((H,B),(G,O),C)	−2626.486	12.074	0.0050	0.0150	0.0013	0.0620
2	((H,B),(C,O),G)	−2627.563	13.150	0.0150	0.0190	0.0019	0.0100
3	((H,B),(C,G),O)	−2627.563	13.150	0.0150	0.0190	0.0019	0.0068
4	((H,C),(G,O),B)	−2621.668	7.256	0.0490	0.1560	0.0270	0.0414
5	((H,C),(B,O),G)	−2614.413	0.000	0.8390	1.0000	0.7224	0.9490
6	((H,C),(B,G),O)	−2621.668	7.256	0.0500	0.1570	0.0270	0.0399
7	((H,G),(C,O),B)	−2627.071	12.659	0.0220	0.0270	0.0040	0.0449
8	((H,G),(B,O),C)	−2618.205	3.793	0.1610	0.4250	0.1187	0.2531
9	((H,G),(B,C),O)	−2627.051	12.639	0.0220	0.0260	0.0043	0.0512
10	((H,O),(C,G),B)	−2627.590	13.177	0.0130	0.0160	0.0017	0.0193
11	((H,O),(B,C),G)	−2627.441	13.029	0.0170	0.0210	0.0025	0.0516
12	((H,O),(B,G),C)	−2626.874	12.461	0.0080	0.0140	0.0010	0.0174
13	((B,G),(C,O),H)	−2626.874	12.461	0.0080	0.0140	0.0010	0.0150
14	((C,G),(B,O),H)	−2618.401	3.989	0.1470	0.4090	0.0833	0.0536
15	((O,G),(B,C),H)	−2626.316	11.904	0.0070	0.0160	0.0019	0.0763

SH, ELW, AU : tests multiples ; KH : test simple entre τ_5 et chacune des topologies τ_j

Plan

Historique

- Théorème de Bayes établi au XVIII^{e} siècle :
  - Utilisation courante en probabilités.
- Introduction récente en phylogénie moléculaire :
  - Yang et Rannala (1996).
- Détermination analytique des probabilités postérieures fréquemment impossible :
  - Utilisation d'approximations numériques.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

Théorème de Bayes

- Une définition classique des probabilités conditionnelles est que :

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B) = \mathbb{P}(A)\mathbb{P}(B|A)$$

- En divisant les deux termes de l'équation précédente par $\mathbb{P}(B)$ on obtient la formulation la plus simple du théorème de Bayes, soit :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}$$

avec :

- $\mathbb{P}(A|B)$, la probabilité *a posteriori* (ou postérieure) de A sachant B.
- $\mathbb{P}(A)$, la probabilité *a priori* de A.
- $\mathbb{P}(B|A)$, la *vraisemblance* de A.
- $\mathbb{P}(B)$, la probabilité *marginale* de B ou *constante de normalisation*.

Généralisation

- Étant donné que :

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(\bar{A} \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(\bar{A})\mathbb{P}(B|\bar{A})$$

on déduit :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(\bar{A})\mathbb{P}(B|\bar{A})}$$

- Ce qui peut se généraliser sous la forme :

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B|A_j)}$$

pour tout élément du s.c.e. $\{A_i\}$, avec i un des éléments de l'ensemble des valeurs possibles de j.

Un exemple classique

- Quelle est la probabilité d'avoir des *faux positifs* lors d'un test de diagnostic ?
- Soit un test de dépistage d'une maladie quelconque :
  - Si un patient a contracté la maladie, le test est positif dans 99% des cas.
  - Si un patient est sain, le test est négatif dans 95% des cas.
  - On estime que la fréquence de la maladie dans la population est de 1‰.
- Quelle est la probabilité qu'un individu testé positif soit effectivement atteint ?

Résolution

- Dans cet exemple, la probabilité *a priori* est égale à la fréquence de la maladie dans la population, soit $\mathbb{P}(A) = 0.001$:
  - On en déduit $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A) = 0.999$.
- Par ailleurs, la probabilité que le test soit positif si le patient est malade est $\mathbb{P}(B|A) = 0.99$.
- Enfin, la probabilité que le test soit négatif si le patient est sain est $\mathbb{P}(\bar{B}|\bar{A}) = 0.95$:
  - On en déduit $\mathbb{P}(B|\bar{A}) = 1 - \mathbb{P}(\bar{B}|\bar{A}) = 0.05$.
- On en déduit la probabilité $\mathbb{P}(A|B)$ qu'un individu soit malade si le test est positif :

$$\mathbb{P}(A|B) = \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.999 \times 0.05} \simeq 0.019$$

Remarques sur le résultat

- Bien que le test précédent soit apparemment précis, la probabilité d'avoir des faux positifs est très importante (98.1%) :
  - Problème lié au fait que la probabilité *a priori* est faible.
  - Cas fréquent pour les tests de diagnostic :
    - Utilisation de plusieurs tests réalisés de façon séquentielle.
- Dans cet exemple, détermination de l'*a priori* à partir de la fréquence de la pathologie dans la population :
  - L'utilisation du théorème de Bayes ne souffre pas de discussion.
- Dans de nombreux cas, les probabilités *a priori* ne peuvent pas être facilement estimées :
  - Utilisation de valeurs représentant l'appréciation *subjective* de la personne effectuant l'analyse.

Notation en statistiques

- En statistiques, le s.c.e. $\{A_i\}$ correspond à un ensemble d'hypothèses, alors que B correspond aux données observées.
- Dans ce cas, écriture du théorème de Bayes sous la forme :

$$\mathbb{P}(H_i|D) = \frac{\mathbb{P}(H_i)\mathbb{P}(D|H_i)}{\sum_j \mathbb{P}(H_j)\mathbb{P}(D|H_j)}$$

avec $\mathbb{P}(H_i|D)$, la probabilité conditionnelle d'une hypothèse H_i sous les données D.

- Les différentes hypothèses pouvant correspondre à différentes valeurs pour un paramètre θ, avec $H_1 : \theta = \theta_1$, $H_2 : \theta = \theta_2$, etc.
- Dans le cas où le modèle utilisé comprend plus d'un paramètre, θ correspond alors au vecteur $\boldsymbol{\theta}$ des dits paramètres.

Données continues

- Expression sous la forme de fonctions de densités quand les hypothèses concernent des paramètres *continus* :

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{f(\mathbf{x})} = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{\int f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}$$

- La constante de normalisation $f(\mathbf{x})$ est obtenue en intégrant la vraisemblance sur la distribution *a priori* de $\boldsymbol{\theta}$:
  - Permet d'avoir $\int f(\boldsymbol{\theta}|\mathbf{x}) = 1$.
  - Si $\boldsymbol{\theta}$ correspond à un vecteur comprenant de nombreux paramètres :
    - Pas de solution analytique au calcul de cette intégrale.
    - Calcul de la probabilité postérieure au moyen d'approximations numériques telles que les *Chaînes de Markov avec technique de Monte-Carlo* (MCMC).

Interprétation des résultats

- Le résultat d'une analyse statistique bayésienne est représenté par la distribution des probabilités postérieures.
- Utilisation de valeurs ponctuelles pour faciliter l'interprétation :
  - Moyenne :

$$\mathbb{E}(\boldsymbol{\theta}|\mathbf{x}) = \int \boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta}$$

  - Médiane.
  - Maximum *a posteriori* :
    - Conceptuellement similaire au maximum de vraisemblance.
- Détermination d'un intervalle de *crédibilité* $[a, b]$ au seuil α tel que :

$$\int_a^b f(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta} = 1 - \alpha$$

Distributions *a priori*

- Conjuguées :
  - Un *a priori* est dit conjugué si $f(\boldsymbol{\theta})$ et $f(\boldsymbol{\theta}|\mathbf{x})$ appartiennent à la même famille de distributions.
  - Permettent de simplifier les calculs (pas de résolution d'intégrales complexes).
- Non informatives ou vagues :
  - $f(\boldsymbol{\theta})$ est non informative si son impact sur $f(\boldsymbol{\theta}|\mathbf{x})$ est faible :
    - Prédominance de la vraisemblance.
  - Utilisées quand aucune information préalable n'est disponible sur les variations du paramètre.
- Informatives :
  - $f(\boldsymbol{\theta})$ est informative si son impact sur $f(\boldsymbol{\theta}|\mathbf{x})$ est fort.
  - Cas de l'analyse bayésienne séquentielle :
    - *A posteriori* d'une étude précédente utilisé comme *a priori* pour l'étude courante.

Critiques de l'*a priori*

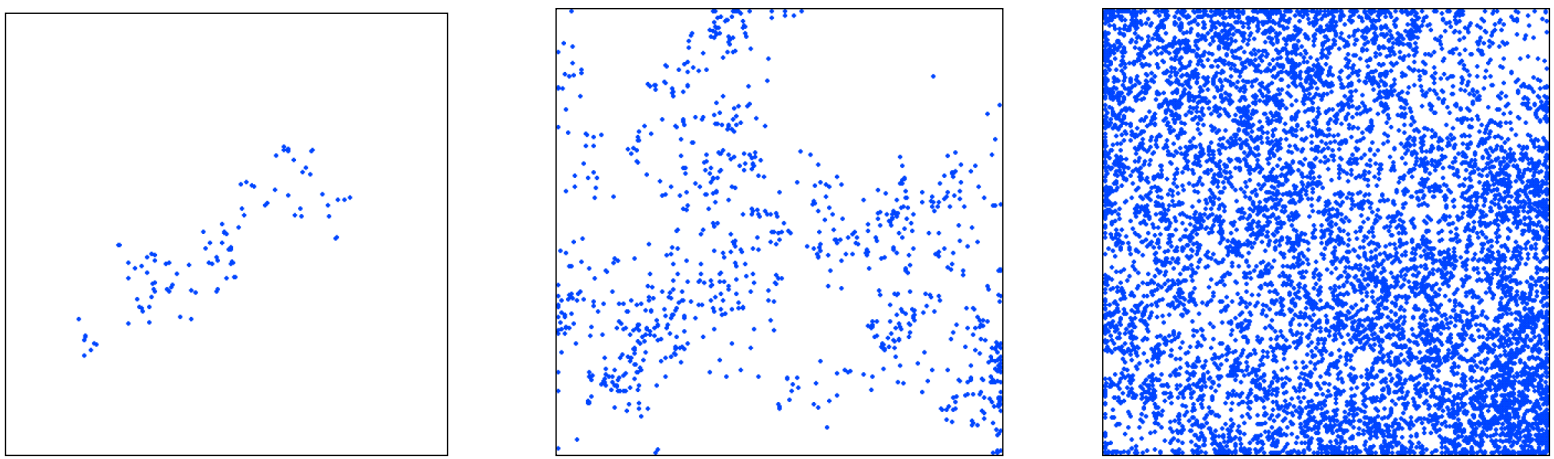
- Depuis le XVIII^{e} siècle, les critiques du bayésien portent essentiellement sur l'*a priori*.
- Résultats différents en fonction d'un *a priori* donné :
  - Rejet par les statisticiens « classiques » de la notion de probabilité subjective.
- Existence d'une école « objective » prônant l'utilisation d'*a priori* les moins informatifs possibles :
  - Distributions uniformes.
  - Loi *a priori* de Jeffreys (1961).
  - Loi de référence de Bernardo (1979).

Principe des MCMC

- En analyse bayésienne, impossibilité de déterminer la constante de normalisation si le nombre de paramètres est élevé :
  - Impossibilité de calculer directement la probabilité postérieure.
- Utilisation d'une chaîne de Markov suivant une marche guidée dans l'espace multidimensionnel des paramètres :
  - À la stationnarité, convergence vers les valeurs attendues des probabilités postérieures.

Analogie du randonneur

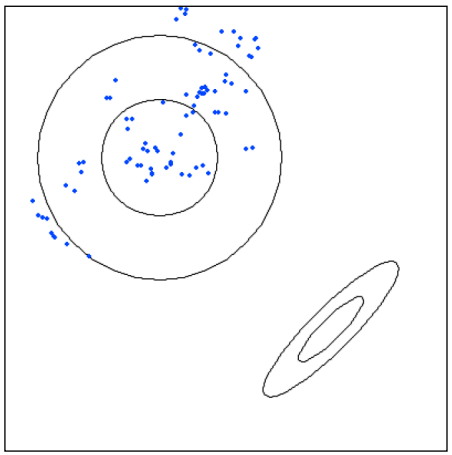
- Soit un randonneur se déplaçant sur une surface plane délimitée en faisant des pas de longueur variable :
  - Amplitude maximale fixée au préalable.
  - Chaque pas est effectué en choisissant aléatoirement une direction quelconque.
  - Rebond si un pas conduit à l'extérieur.
- Au bout d'un certain temps, exploration de l'intégralité de la surface :



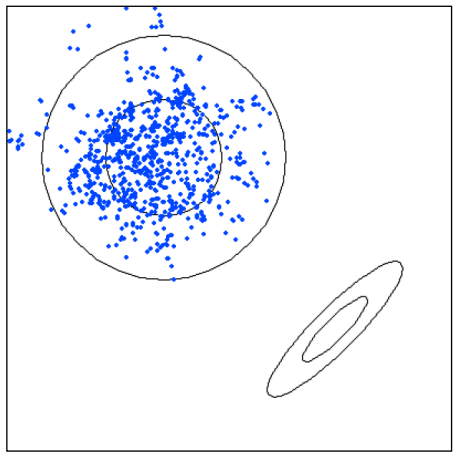
100 pas 1000 pas 10000 pas

Exploration de reliefs

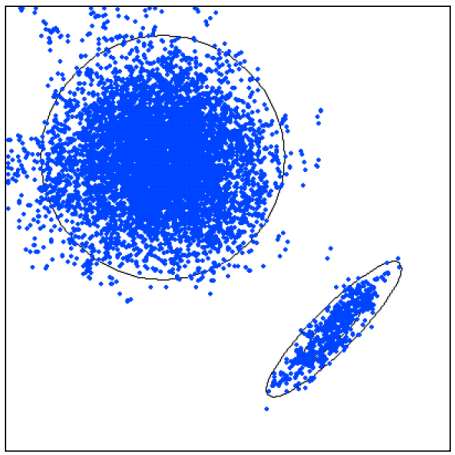
- Introduction de deux règles supplémentaires :
  - Si la direction prise par le randonneur le conduit vers une position plus élevée, il le fait toujours.
  - Si au contraire cette direction est descendante, possibilité de choix :
    - Calcul de $r = h^*/h$, avec h^* la hauteur atteinte en cas de descente et h la hauteur actuelle.
    - Tirage de $u \sim \mathcal{U}(0, 1)$.
    - Si $u < r$, le randonneur descend, sinon il reste où il est.
- Visite préférentielle des points situés en altitude :



100 pas



1000 pas



10000 pas

Problèmes rencontrés

- Nécessité d'éliminer les premiers pas – qui constituent ce que l'on appelle communément la *zone d'approche* ou *burn-in* :
  - Démarrage du trajet en un point sélectionné aléatoirement, point pouvant être situé à une distance importante des reliefs.
- Évitement des maxima locaux :
  - Nécessité d'avoir un nombre de pas suffisamment élevé :
    - Pas toujours suffisant si les pics sont éloignés les uns des autres.
  - Lancement de plusieurs chaînes ayant des points de départ différents :
    - Poursuite de l'exploration jusqu'à convergence des résultats entre les différentes chaînes.

Algorithme de Metropolis-Hastings

1. Soit $\boldsymbol{\theta}_i$, le vecteur des paramètres caractérisant l'état de la chaîne de Markov au temps i.
2. Soit $\boldsymbol{\theta}^*$ le vecteur des paramètres caractérisant un état *candidat* pour constituer le maillon suivant de la chaîne.
3. Calcul de la *probabilité d'acceptation* r, telle que :

$$r = \min\left[1, \frac{f(\boldsymbol{\theta}^*|\mathbf{x})}{f(\boldsymbol{\theta}_i|\mathbf{x})}\right] = \min\left[1, \frac{f(\boldsymbol{\theta}^*)f(\mathbf{x}|\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_i)f(\mathbf{x}|\boldsymbol{\theta}_i)}\right]$$

4. Si $r = 1$, alors $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$.
5. Si $r < 1$, tirage de $u \sim \mathcal{U}(0,1)$:
   - Si $u < r$ alors $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$, sinon $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$.
6. Retour à l'étape 1.

Caractéristiques

- Le calcul de r n'implique pas de connaître $f(\mathbf{x})$.
- Initialisation avec un ensemble de paramètres $\boldsymbol{\theta}$ choisis aléatoirement.
- La construction de $\boldsymbol{\theta}^*$ se fait en faisant varier de façon aléatoire les paramètres :
  - Utilisation d'algorithmes générant ce que l'on appelle des *propositions* :
    - Distributions uniformes de type $\mathcal{U}(-w/2, w/2)$, avec w l'amplitude maximale autorisée pour la variation des paramètres.
    - Distributions normales de type $\mathcal{N}(\mu, \sigma^2)$.
- La séquence des états visités forme une chaîne de Markov :
  - Estimation de la probabilité postérieure par la fréquence à laquelle les états sont visités une fois la stationnarité atteinte.

Fréquence d'acceptation

- Proportion du nombre de propositions acceptées dans la chaîne.
- Ne doit être ni trop grande ni trop petite.
- Valeurs optimales :
  - $\approx 50\%$ si $\boldsymbol{\theta}$ ne comprend qu'un seul paramètre.
  - $\approx 26\%$ si $\boldsymbol{\theta}$ comprend plusieurs paramètres.
- Valeurs recommandées :
  - 20-70% si $\boldsymbol{\theta}$ ne comprend qu'un seul paramètre.
  - 15-40% si $\boldsymbol{\theta}$ comprend plusieurs paramètres.

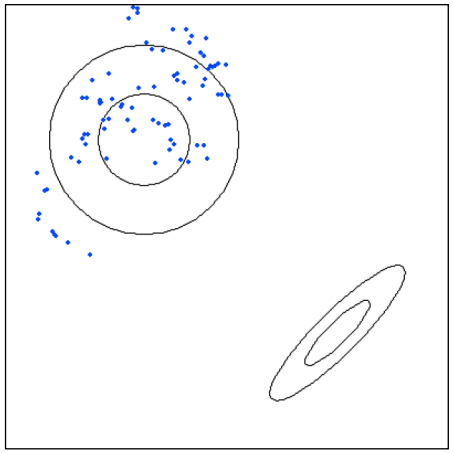
Couplage de Metropolis des MCMC

- Piégeage possible de la chaîne en cas de maximum local.
- Utilisation de plusieurs chaînes au lieu d'une :
  - Couplage de Metropolis des MCMC (MCMCMC ou MC^3).
  - Parmi toutes les chaînes lancées seules les chaînes dites « froides » (faible amplitude des pas) ont besoin de converger :
    - Utilisation de chaînes « chaudes » pour permettre une exploration plus vaste de l'espace des paramètres.
  - Tests à intervalles réguliers pour faire passer une chaîne froide dans une région explorée par une des chaînes chaudes :

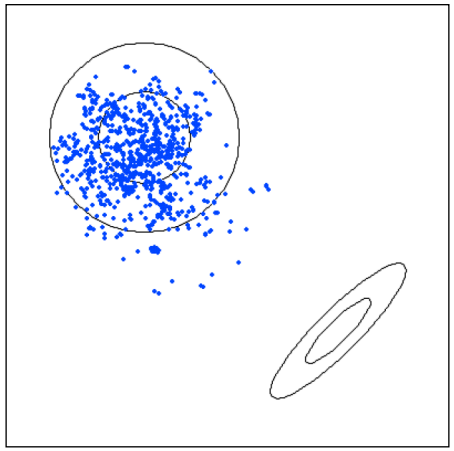
$$r = \min\left[1, \frac{\pi_i(\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_i)}{\pi_i(\boldsymbol{\theta}_i)\pi_j(\boldsymbol{\theta}_j)}\right]$$

où i et j correspondent aux états de deux chaînes de Markov pour lesquelles la possibilité d'échange est testée.

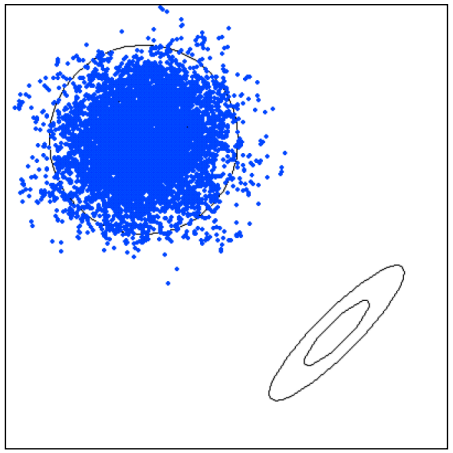
Application au problème du randonneur



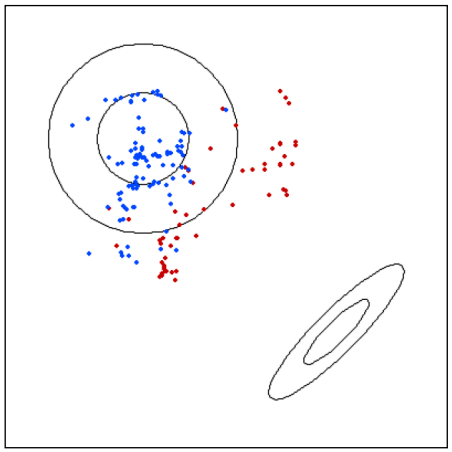
100 pas

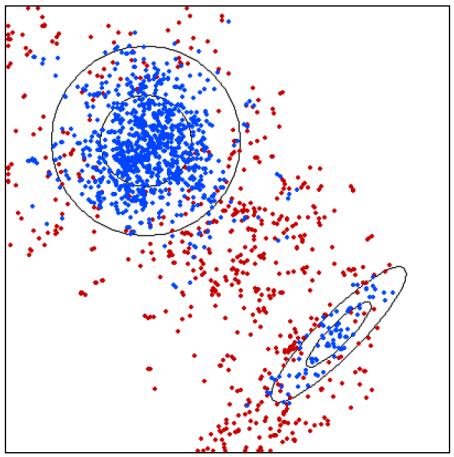


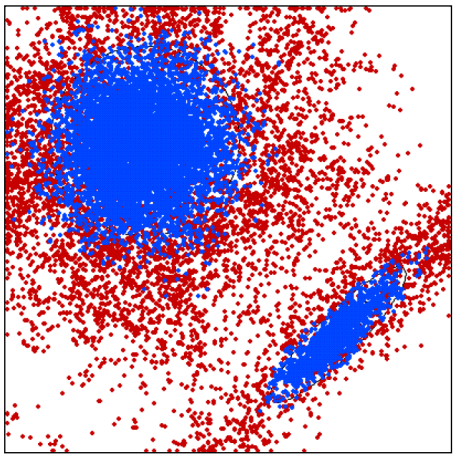
1000 pas



10000 pas







Détermination de la convergence

- Quand faut-il interrompre une MCMC ?
  - A-t-on atteint la distribution stationnaire de la chaîne ?
- Outils disponibles :
  - Inspection visuelle du graphe montrant les déplacements dans l'espace des paramètres.
  - Étude de la variation des valeurs de vraisemblance :
    - Pas de tendances particulières attendues à la stationnarité.
  - Mesure de l'autocorrélation des valeurs successives des paramètres :
    - Absence d'autocorrélation si convergence.
  - Tests statistiques :
    - Test de Gelman et Rubin (1992), ou *Potential Scale Reduction Factor* (PSRF) dans MrBayes.

Probabilité *a priori*

- Estimation par approche bayésienne de la distance évolutive entre deux séquences d'ADN sous le modèle de Jukes et Cantor.
- Calcul de la probabilité *a priori* :
  - Choix d'une distribution exponentielle :

$$f(t) = \frac{1}{\mu} \mathrm{e}^{-t/\mu}$$

    avec μ la moyenne de cette distribution et t la distance évolutive :
    - La probabilité d'obtenir des distances importantes tend rapidement vers 0.
  - D'autres choix sont possibles :
    - Distribution uniforme.

Vraisemblance

- Le calcul de la divergence observée entre deux séquences au moyen du modèle de Jukes et Cantor est donnée par (*cf.* Diapo. 56) :

$$p = 3p_{ij}(t) = \frac{3}{4} - \frac{3}{4}\mathrm{e}^{-4t/3} \quad (i \neq j)$$

- 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 t est donnée par la distribution binomiale $\mathcal{B}(\ell, p)$ telle que :

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

Probabilité postérieure

- Probabilité postérieure, sans la constante de normalisation :

$$\begin{aligned} f(t|p) &\propto f(t)f(p|t) \\ &\propto \frac{1}{\mu}\mathrm{e}^{-t/\mu}\left(\frac{3}{4}-\frac{3}{4}\mathrm{e}^{-4t/3}\right)^{n}\left(\frac{1}{4}+\frac{3}{4}\mathrm{e}^{-4t/3}\right)^{\ell-n} \end{aligned}$$

  Le coefficient binomial étant lui aussi une constante, il peut être omis de cette expression.

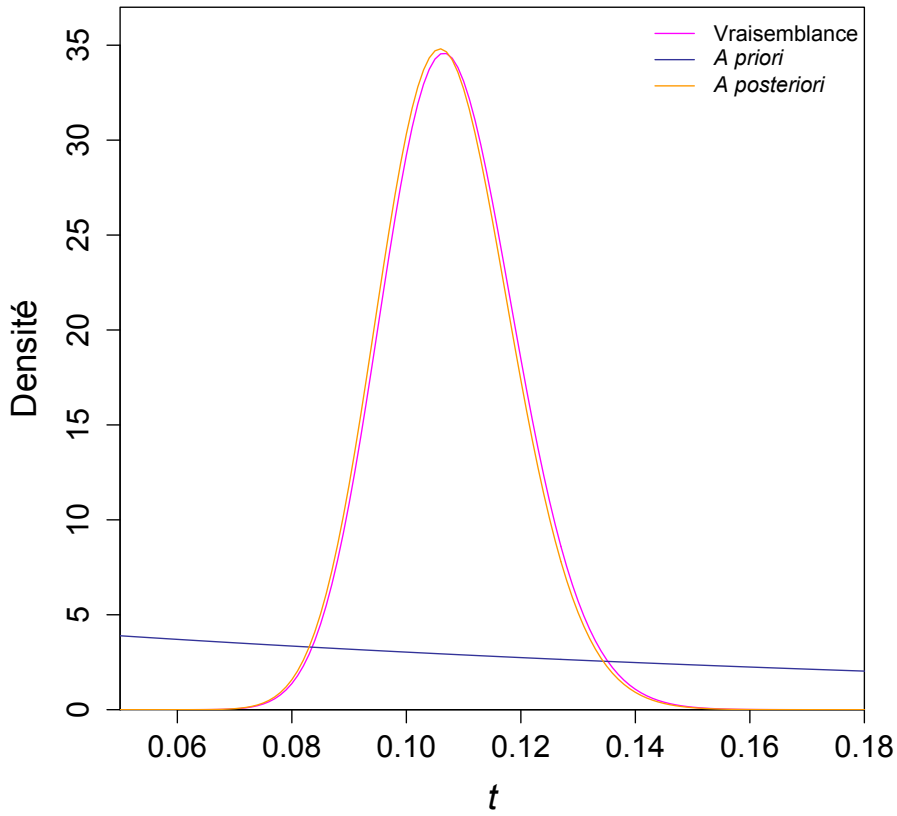
- Valeur de la constante de normalisation donnée par :

$$f(p) = \int_0^\infty f(t)f(p|t)\mathrm{d}t$$

  Solution analytique ou intégration numérique.

Application numérique

- Paire Homme-Gorille du jeu de données de Brown *et al.* (1982) :
  - $\ell = 896$
  - $n = 89$
- Moyenne de la distribution *a priori* fixée à $\mu = 0.2$.
- Estimation au maximum de vraisemblance :
  - $t \simeq 0.1066$
- Estimation bayésienne via la moyenne :
  - $\mathbb{E}(t|p) \simeq 0.1072$



Approximation par MCMC

- Calcul de la probabilité d'acceptation :

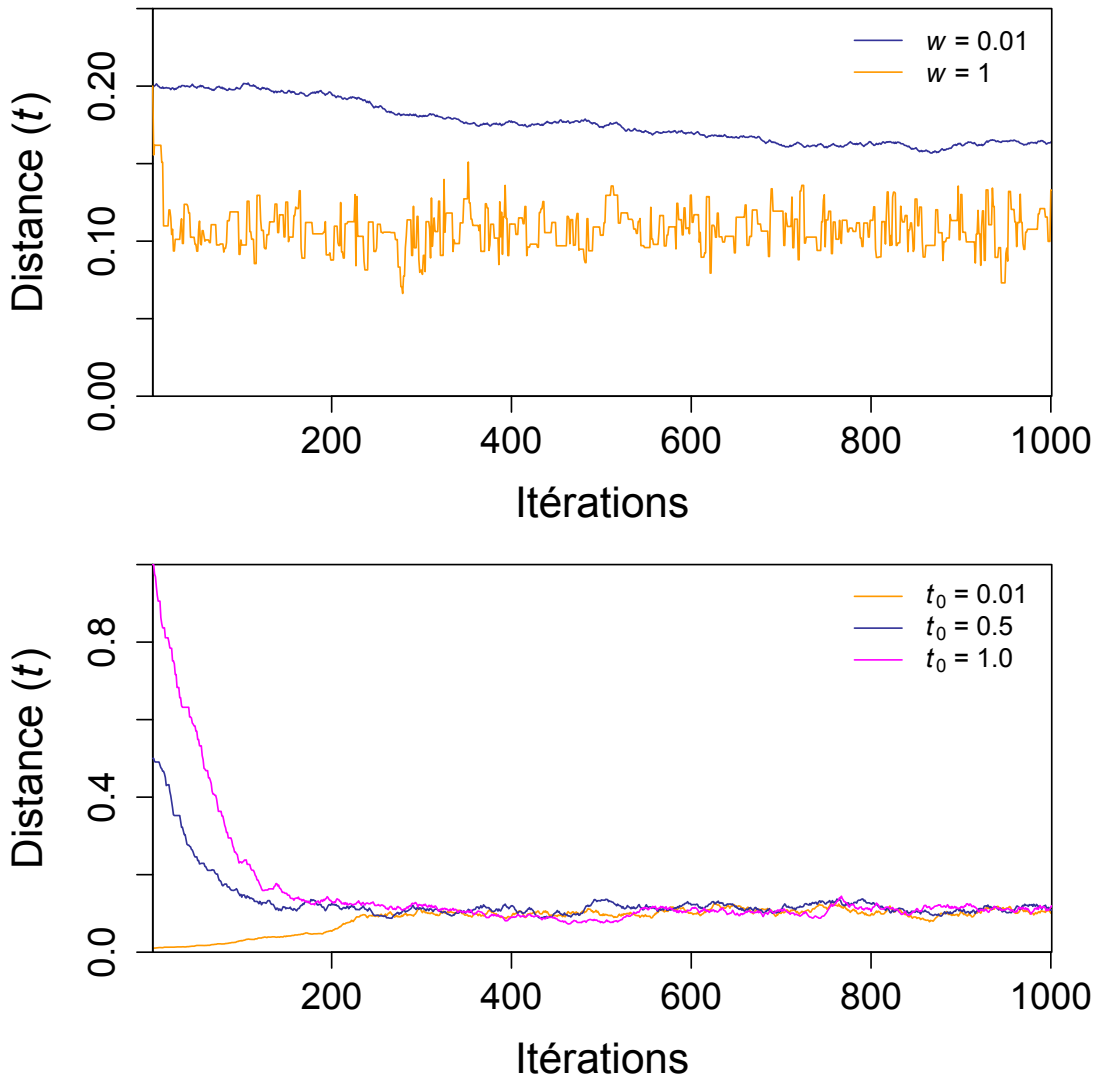
$$r = \min\left[1, \frac{f(t^*)f(p|t^*)}{f(t_i)f(p|t_i)}\right]$$

- Choix des propositions pour t :
  - Distribution uniforme, centrée sur la valeur actuelle et ayant une largeur égale à w :

$$t^* = |t_i + u|, \text{ avec } u \sim \mathcal{U}(-w/2, w/2)$$

- Choix de différentes valeurs pour l'amplitude (w) et la distance (t_0) utilisées pour initialiser la chaîne de Markov :
  - Valeurs variables pour w (0.01 et 1) et valeur fixe pour t_0 (0.2).
  - Valeur fixe pour w (0.1) et valeurs variables pour t_0 (0.01, 0.5 et 1).

Convergence des chaînes



Estimations de la distance

- Paramètres choisis : $\mu = 0.2$, $w = 0.1$ et $t_0 = 0.5$.
- Élimination de la zone d'approche (400 premières itérations).
- Échantillonnage de 1000 itérations prélevées à intervalles réguliers dans une chaîne :
  - Utilisation de la moyenne des valeurs pour l'estimation.
- Estimations obtenues après :
  - 1400 itérations : $\mathbb{E}(t|p) = 0.1073 \pm 5.18 \times 10^{-4}$
  - 10000 itérations : $\mathbb{E}(t|p) = 0.1072 \pm 7.03 \times 10^{-4}$
  - 100000 itérations : $\mathbb{E}(t|p) = 0.1071 \pm 7.23 \times 10^{-4}$

  avec, dans chaque cas, un intervalle de crédibilité à 95%.
- Variations stochastiques autour de la valeur obtenue par calcul direct.

Notations pour la phylogénie

- En phylogénie moléculaire, les données sont représentées par un ensemble de séquences alignées S.
- Par ailleurs, le vecteur des paramètres est $\boldsymbol{\theta} = (\tau, \mathbf{t}, \boldsymbol{\vartheta}, \alpha)$, avec :
  - τ la topologie de l'arbre.
  - $\mathbf{t}$ le vecteur des longueurs de branches.
  - $\boldsymbol{\vartheta}$ le vecteur des paramètres du modèle d'évolution utilisé.
  - α le paramètre de forme de la loi Gamma, le cas échéant.
- Le formule permettant de déterminer la probabilité postérieure est donc égale à :

$$f(\tau, \mathbf{t}, \boldsymbol{\vartheta}, \alpha | S) = \frac{f(\tau, \mathbf{t}, \boldsymbol{\vartheta}, \alpha) f(S | \tau, \mathbf{t}, \boldsymbol{\vartheta}, \alpha)}{f(S)}$$

avec :

$$f(S) = \sum_{\tau} \int_{\mathbf{t}} \int_{\boldsymbol{\vartheta}} \int_{\alpha} f(S | \tau, \mathbf{t}, \boldsymbol{\vartheta}, \alpha) f(\mathbf{t}) f(\boldsymbol{\vartheta}) f(\alpha) \mathrm{d}\mathbf{t} \mathrm{d}\boldsymbol{\vartheta} \mathrm{d}\alpha$$

Choix possibles pour les *a priori*

- Topologies :
  - Distribution uniforme $\mathcal{U}(N)$.
- Longueurs des branches :
  - Distribution uniforme $\mathcal{U}(0, 10)$.
  - Distribution exponentielle $\mathcal{E}(0.1)$.
- Paramètres du modèle d'évolution :
  - Distributions de Dirichlet plates $\mathcal{D}(1, 1, 1, 1)$ pour les échangeabilités et les fréquences à l'équilibre.
- Paramètre α de la loi Gamma :
  - Distribution exponentielle $\mathcal{E}(1)$.

Facteur de Bayes

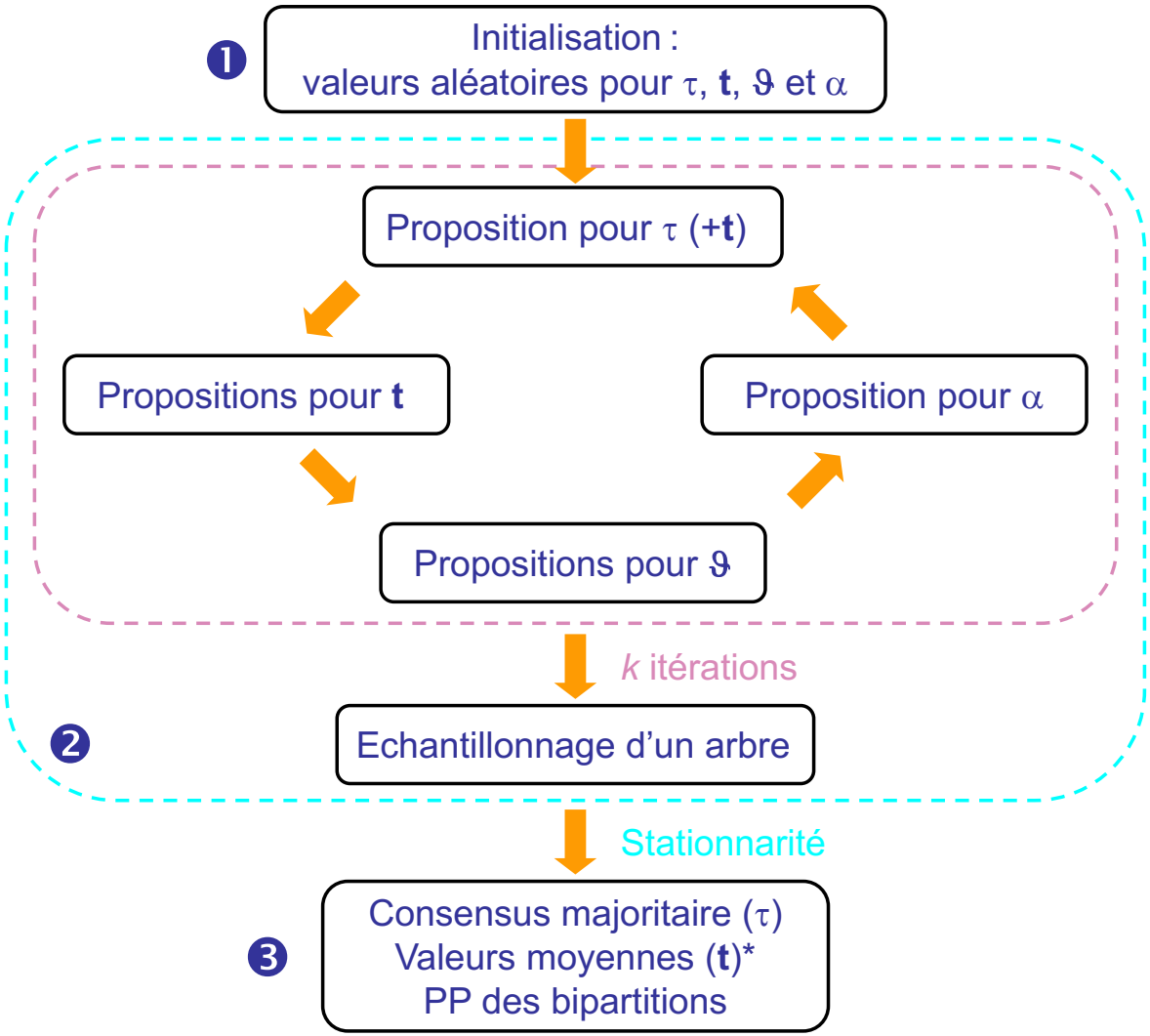
- Défini comme étant le rapport des *vraisemblances marginales* associées aux modèles M_0 et M_1 comparés, soit :

$$\mathrm{BF}_{10} = \frac{f(\mathbf{x}|M_1)}{f(\mathbf{x}|M_0)} = \frac{\int f(\boldsymbol{\vartheta}_1|M_1) f(\mathbf{x}|\boldsymbol{\vartheta}_1, M_1) \mathrm{d}\boldsymbol{\vartheta}_1}{\int f(\boldsymbol{\vartheta}_0|M_0) f(\mathbf{x}|\boldsymbol{\vartheta}_0, M_0) \mathrm{d}\boldsymbol{\vartheta}_0}$$

- Si $H_0 = M_0$, alors interprétation en utilisant l'échelle de Kass et Raftery (1995) :

log(BF)	BF	Évidence
< 0	< 1	Négative
$0 - 0.5$	$1 - 3.2$	Faible
$0.5 - 1$	$3.2 - 10$	Substancielle
$1 - 2$	$10 - 100$	Forte
> 2	> 100	Décisive

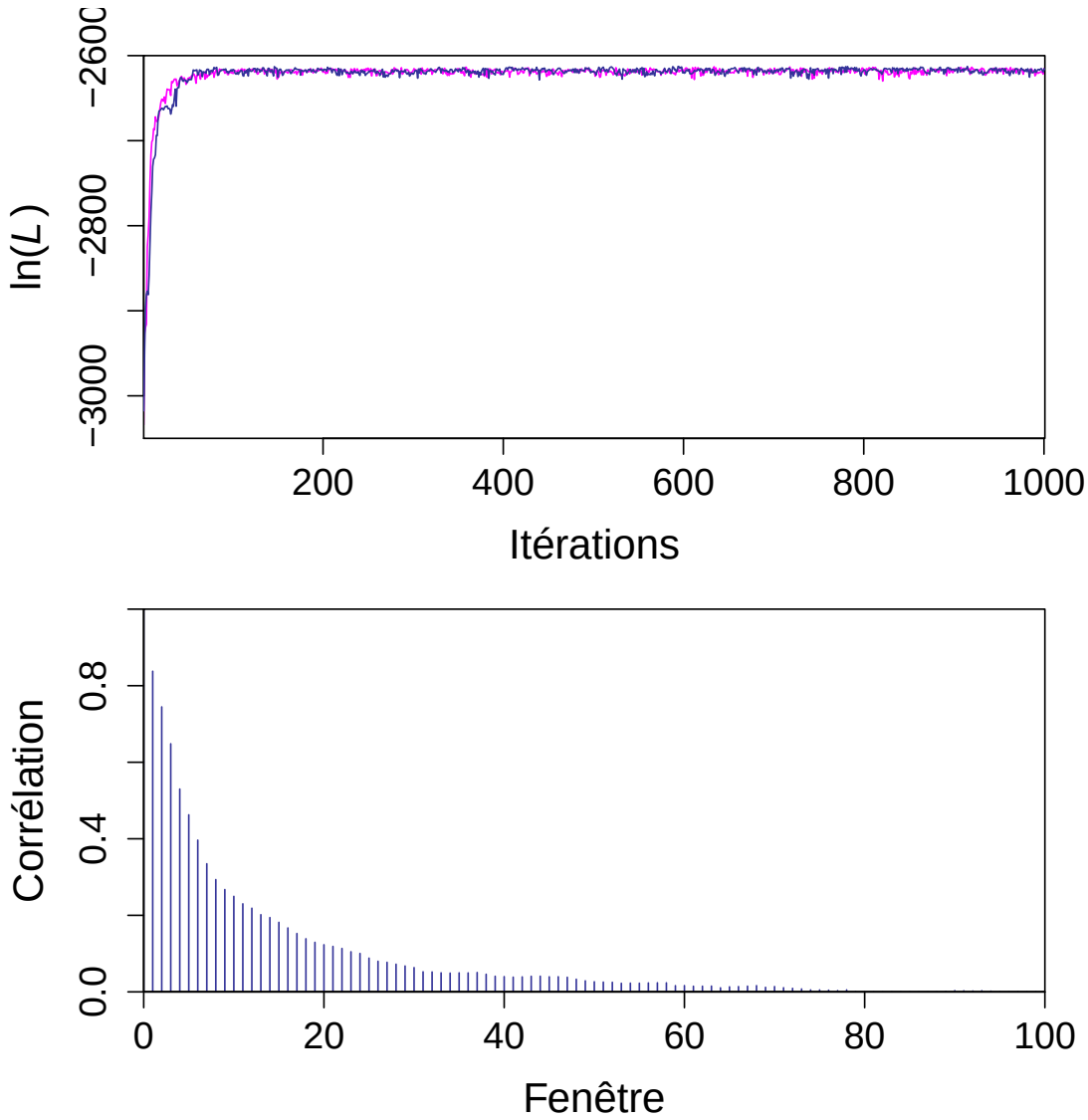
Procédure générale



Phylogénie des Hominoïdes

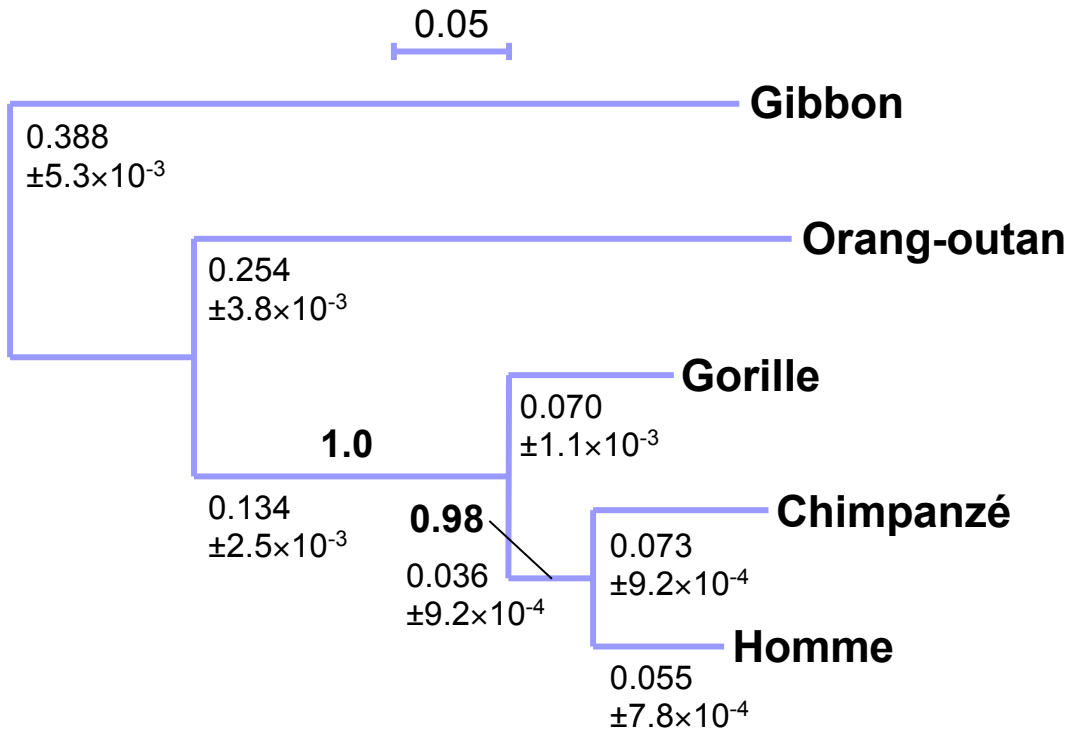
- Sélection du modèle HKY+Γ après un test BIC.
- Utilisation de MrBayes pour reconstruire la phylogénie :
  - Valeurs par défaut des probabilités *a priori*.
  - Deux chaînes froides partant de points de départ différents.
  - Trois chaînes chaudes lancées en parallèle de chaque chaîne froide.
  - Test de Gelman et Rubin pour déterminer si convergence.
  - Arrêt après 10000 itérations et fréquence d'échantillonnage de 1/10 :
    - Jeu de données de petite taille.

Convergence des chaînes



Arbre obtenu

- Construction par consensus majoritaire à 50% sur les itérations échantillonnées hors *burn-in*.
- Racinement avec la séquence du Gibbon.
- Longueurs des branches avec intervalles de crédibilité à 95%.

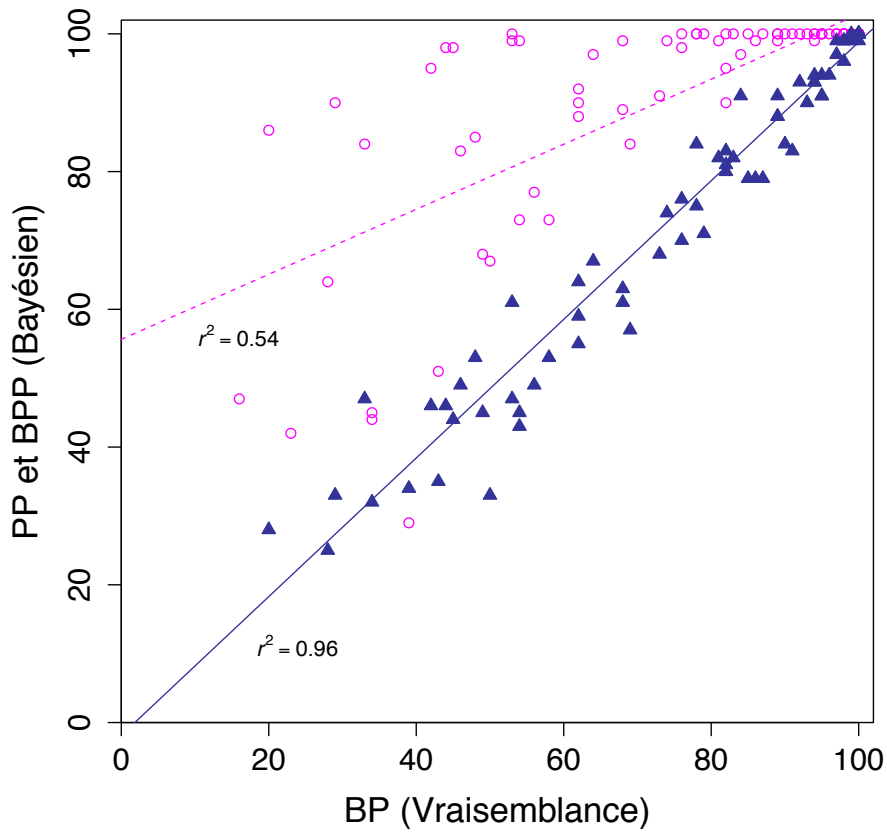


Avantages et limitations

- Meilleur comportement que le maximum de vraisemblance avec des modèles comprenant de nombreux paramètres :
  - Intégration des paramètres de nuisance.
- Temps de calcul biens plus longs :
  - Avec les MC^3, de nombreuses chaînes sont lancées en parallèle.
  - Nécessité d'atteindre la distribution stationnaire pour les chaînes froides :
    - Diminution du nombre d'itérations pour raccourcir les temps de calcul.
- Pas de nécessité d'effectuer du rééchantillonnage de type *bootstrap* :
  - Utilisation des valeurs de probabilités postérieures des clades :
    - Valeurs directement interprétables en termes de probabilités.

Bootstrap et probabilités postérieures

- Construction de six phylogénies (Douady *et al.*, 2003) :
  - Vraisemblance et bayésien.
- Comparaison entre valeurs de *bootstrap* (BP) et :
  - Probabilités postérieures (PP) des clades.
  - *Bootstrap* des probabilités postérieures (BPP).
- Valeurs des PP systématiquement plus élevées que celles des BP.



Plan

Notations mathématiques

- A : objet non mathématique (*e.g.*, nucléotide, UTO).
- a : définition d'une variable.
- e : constante ou opérateur.
- A : définition d'une variable ou d'un ensemble (hors ensembles numériques).
- $\mathbf{A}$: matrice ou vecteur (ligne ou colonne).
- $\mathbf{a}$: vecteur (ligne ou colonne).
- $\mathcal{B}$: distribution (*e.g.*, Binomiale, Gamma).
- $\mathbb{P}$: terme mathématique usuel (*e.g.*, probabilité, variance) ou ensemble numérique (*e.g.*, entiers naturels, réels).

Codes IUPAC pour les nucléotides

Code	Signification	Compl.
A	A	T/U
C	C	G
G	G	C
T/U	T/U	A
M	A/C	K
R	A/G	Y
S	C/G	S
W	A/T/U	W
Y	C/T/U	R
K	G/T/U	M
V	A/C/G	B
H	A/C/T/U	D
D	A/G/T/U	H
B	C/G/T/U	V
N	A/C/G/T/U	N

Matrices et vecteurs

- Une matrice $m \times n$ est un tableau à m lignes et n colonnes :
  - On note $\mathbf{A} = (a_{ij})$ $(1 \leq i \leq m$ et $1 \leq j \leq n)$ la matrice dont a_{ij} est l'élément de la $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne :

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

- Un vecteur colonne est une matrice $m \times 1$ tandis qu'un vecteur ligne est une matrice $1 \times n$:
  - On note $\mathbf{v} = (v_i)$ $(1 \leq i \leq m)$ un vecteur colonne et $\mathbf{v} = (v_j)$ $(1 \leq j \leq n)$ un vecteur ligne.

Opérations sur les matrices I

- Transposition d'une matrice :
  - Soit $\mathbf{A} = (a_{ij})$ une matrice $m \times n$.
  - La transposée de $\mathbf{A}$ est une matrice $n \times m$, notée $\mathbf{A}^{\mathrm{T}} = (a_{ij}^{\mathrm{T}})$ et d'élément général $a_{ij}^{\mathrm{T}} = a_{ji}$.
- Somme de deux matrices :
  - Soient $\mathbf{A} = (a_{ij})$ et $\mathbf{B} = (b_{ij})$ deux matrices $m \times n$.
  - La somme de $\mathbf{A}$ et $\mathbf{B}$, notée $\mathbf{A} + \mathbf{B}$, est la matrice $\mathbf{C} = (c_{ij})$, d'élément général $c_{ij} = a_{ij} + b_{ij}$.
- Produit d'une matrice par un scalaire :
  - Soit $\mathbf{A} = (a_{ij})$ une matrice $m \times n$ et soit r un scalaire.
  - La multiplication de $\mathbf{A}$ et de r, notée $r\mathbf{A}$, est la matrice $\mathbf{B} = (b_{ij})$, d'élément général $b_{ij} = ra_{ij}$.

Opérations sur les matrices II

- Produit de deux matrices :
  - Soit $\mathbf{A} = (a_{ij})$ une matrice $m \times n$ et $\mathbf{B} = (b_{ij})$ une matrice $n \times p$.
  - Le produit de $\mathbf{A}$ et $\mathbf{B}$, noté $\mathbf{AB}$, est la matrice $\mathbf{C} = (c_{ij})$, de dimensions $m \times p$ et d'élément général :

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}$$

  La multiplication de $\mathbf{A}$ et $\mathbf{B}$ n'est possible que si le nombre de colonnes de $\mathbf{A}$ est égal au nombre de lignes de $\mathbf{B}$.
  - Si $\mathbf{A}$ et $\mathbf{B}$ sont deux matrices carrées (*cf.* Diapo. 245) alors les produits $\mathbf{AB}$ et $\mathbf{BA}$ ont un sens mais, en général, $\mathbf{AB} \neq \mathbf{BA}$.

Matrices carrées

- Une matrice *carrée* $\mathbf{A} = (a_{ij})$ $(1 \leq i \leq n$ et $1 \leq j \leq n)$ est une matrice ayant le même nombre de lignes et de colonnes :
  - Une matrice carrée $n \times n$ est dite *d'ordre* n.
- Une matrice carrée $\mathbf{A}$ est dite *symétrique* si $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$.
- Une matrice carrée $\mathbf{A}$ est dite *diagonale* si elle vérifie que $a_{ij} = 0,\ \forall i \neq j$:
  - Utilisation de la notation $\mathbf{A} = \mathrm{diag}(a_k)$ $(1 \leq k \leq n)$.
- La matrice *identité* $\mathbf{I}_n$ est la matrice carrée d'ordre n ayant des 1 sur sa diagonale et des 0 partout ailleurs :
  - Elément neutre pour la multiplication : $\mathbf{A}\mathbf{I}_n = \mathbf{I}_n\mathbf{A} = \mathbf{A}$.
- La *trace* d'une matrice carrée $\mathbf{A}$ est égale à la somme de ses éléments diagonaux : $\mathrm{trace}(\mathbf{A}) = \sum_{i=j} a_{ij}$.

Déterminant d'une matrice

- Fonction qui associe à une matrice carrée un nombre réel :
  - Si $\det(\mathbf{A}) = 0$: matrice *singulière*.
  - Si $\det(\mathbf{A}) \neq 0$: matrice *régulière*.
- Nombreuses méthodes de calcul :
  - Formule de Leibnitz.
  - Méthode des cofacteurs.
  - Pivot de Gauss.
- Cas d'une matrice carrée d'ordre 2 :

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \Rightarrow \det(\mathbf{A}) = a_{11}\,a_{22} - a_{21}\,a_{12}$$

Méthode des cofacteurs

- Soit $\mathbf{A}$ une matrice carrée d'ordre n.
- Le *cofacteur* c_{ij} du terme a_{ij} de cette matrice est défini comme :

$$c_{ij} = (-1)^{i+j} \det(\mathbf{A}_{ij})$$

avec $\det(\mathbf{A}_{ij})$ le déterminant de la sous-matrice $\mathbf{A}_{ij}$ obtenue en éliminant la $i^{\text{ème}}$ ligne et la $j^{\text{ème}}$ colonne de $\mathbf{A}$.

- Dans ce cas, le déterminant de $\mathbf{A}$ est égal à :

$$\det(\mathbf{A}) = \sum_{j=1}^{n} a_{ij}\, c_{ij}$$

avec i fixé à l'une des n valeurs possibles.

Inverse d'une matrice

- *L'inverse* d'une matrice carrée $\mathbf{A}$ d'ordre n est la matrice carrée $\mathbf{A}^{-1}$ de même ordre qui vérifie $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$.
- Seules les matrices dont le déterminant est non nul sont inversibles.
- Calcul par la méthode des cofacteurs :
  - Soit $\mathbf{A}$ une matrice carrée d'ordre n et soit $\mathbf{C} = (c_{ij})$ la matrice de l'ensemble des cofacteurs de $\mathbf{A}$, dans ce cas :

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\mathbf{C}^{\mathrm{T}}$$

- Si $\mathbf{A}$ et $\mathbf{B}$ sont deux matrices inversibles de même ordre alors :
  - Le produit $\mathbf{AB}$ est inversible et son inverse est tel que :

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Valeurs propres et vecteurs propres

- Soit $\mathbf{A}$ une matrice carrée d'ordre n, dans ce cas λ_k $(1 \leq k \leq n)$ est une *valeur propre* et $\mathbf{u}_k$ un *vecteur propre* de $\mathbf{A}$ si :

$$\mathbf{A}\mathbf{u}_k = \lambda_k \mathbf{u}_k$$

- Si une matrice d'ordre n admet n valeurs propres distinctes, alors est elle dite *diagonalisable* et peut s'écrire sous la forme :

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

avec $\mathbf{U}$ la matrice contenant en colonnes les vecteurs propres de $\mathbf{A}$ et $\mathbf{\Lambda} = \text{diag}(\lambda_k)$ la matrice diagonale des valeurs propres.

Quelques propriétés

- Toute matrice symétrique est diagonalisable.
- La trace d'une matrice diagonalisable est égale à la somme de ses valeurs propres :

$$\text{trace}(\mathbf{A}) = \sum_{k=1}^{n} \lambda_k = \text{trace}(\mathbf{\Lambda})$$

- Le déterminant d'une matrice diagonalisable est égal au produit de ses valeurs propres :

$$\det(\mathbf{A}) = \prod_{k=1}^{n} \lambda_k = \det(\mathbf{\Lambda})$$

Simplification des calculs I

Soit $\mathbf{A}$ une matrice carrée d'ordre n, diagonalisable :

- Puissance de $\mathbf{A}$:

$$\begin{aligned}\mathbf{A}^m &= \underbrace{\mathbf{A}\mathbf{A}\dots\mathbf{A}}_{m \text{ facteurs}} \\ &= \underbrace{\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}\dots\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}}_{m \text{ facteurs}} \\ &= \mathbf{U}\mathbf{\Lambda}^m\mathbf{U}^{-1}\end{aligned}$$

avec :

$$\mathbf{\Lambda}^m = \begin{pmatrix} \lambda_1^m & 0 & \cdots & 0 \\ 0 & \lambda_2^m & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^m \end{pmatrix}$$

Simplification des calculs II

- Logarithme de $\mathbf{A}$:

$$\ln(\mathbf{A}) = \mathbf{U} \ln(\mathbf{\Lambda}) \mathbf{U}^{-1} \text{ avec } \ln(\mathbf{\Lambda}) = \begin{pmatrix} \ln \lambda_1 & 0 & \cdots & 0 \\ 0 & \ln \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \ln \lambda_n \end{pmatrix}$$

- Exponentielle de $\mathbf{A}$:

$$\mathrm{e}^{\mathbf{A}} = \mathbf{U} \mathrm{e}^{\mathbf{\Lambda}} \mathbf{U}^{-1} \text{ avec } \mathrm{e}^{\mathbf{\Lambda}} = \begin{pmatrix} \mathrm{e}^{\lambda_1} & 0 & \cdots & 0 \\ 0 & \mathrm{e}^{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{e}^{\lambda_n} \end{pmatrix}$$

Variables aléatoires discrètes

- Les valeurs prises par les variables sont discrètes.
- Leur loi est complètement déterminée par $\mathbb{P}(X = x)$ pour tout $x \in \Omega$, avec :

$$\sum_{x \in \Omega} \mathbb{P}(X = x) = 1$$

- La *fonction de répartition* F de X est définie par :

$$F(x) = \mathbb{P}(X \leq x)$$

- La moyenne et la variance sont définies par :

$$\mathbb{E}(X) = \sum_{x \in \Omega} x\, \mathbb{P}(X = x) \text{ et } \mathbb{V}(X) = \sum_{x \in \Omega} [x - \mathbb{E}(X)]^2 \mathbb{P}(X = x)$$

Distribution Binomiale

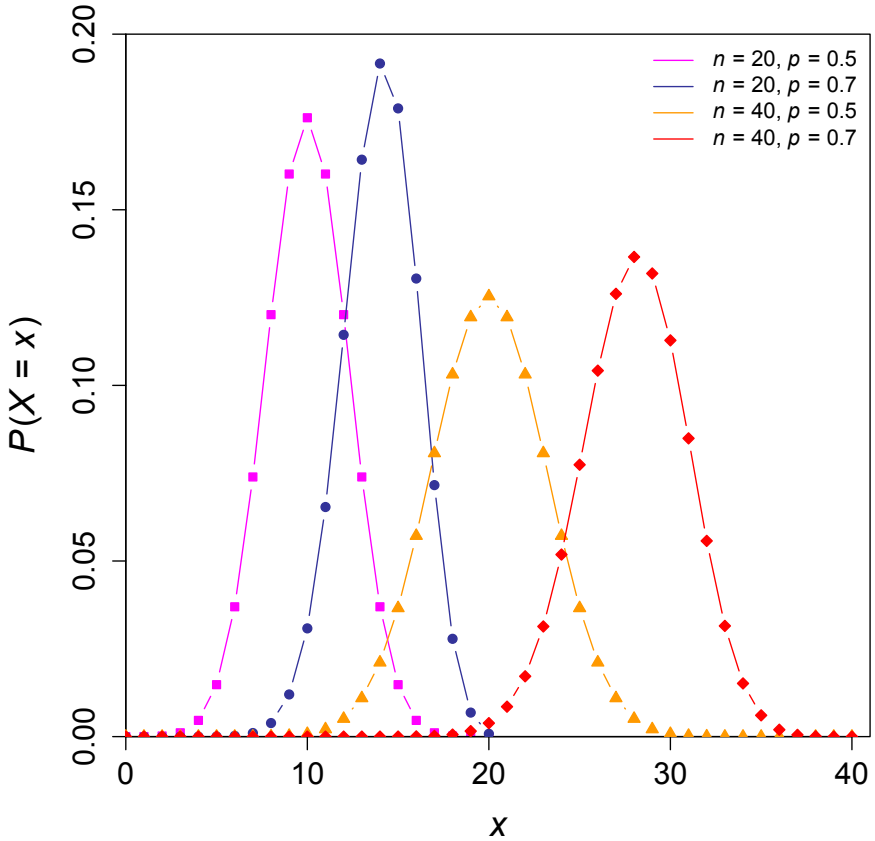
- Écriture sous la forme $\mathcal{B}(n, p)$, avec $n \in \mathbb{N}$ et $p \in [0, 1]$ les deux paramètres de la loi.
- Loi de probabilité :

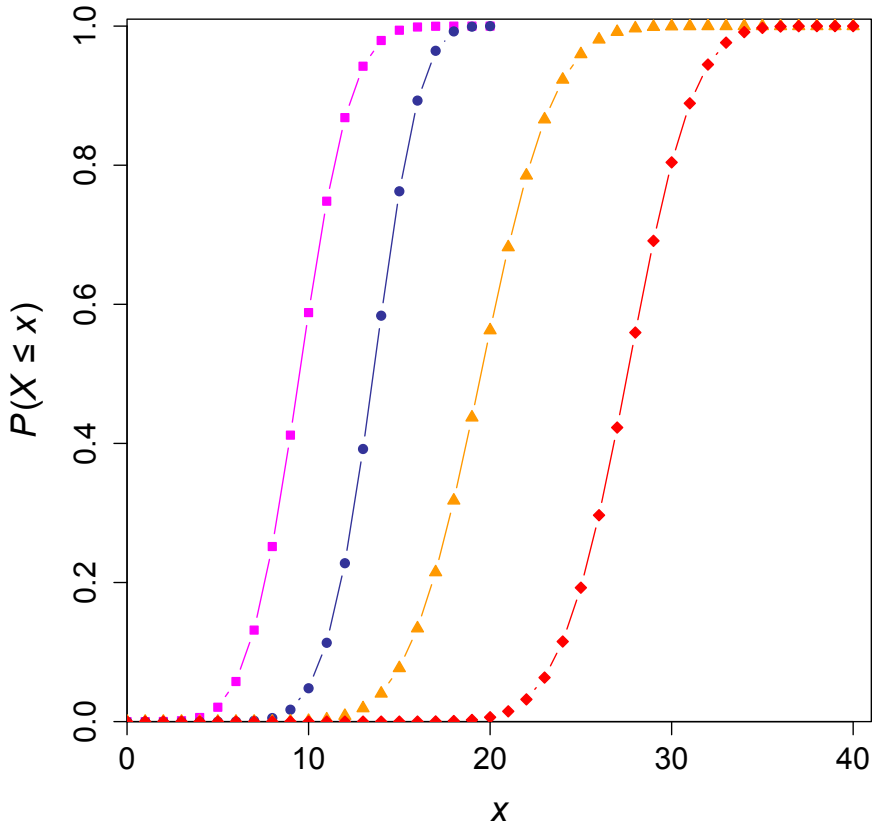
$$\begin{aligned}\mathbb{P}(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}\end{aligned}$$

  avec $\Omega = [0, n]$.
- Moyenne et variance :

$$\mathbb{E}(X) = np \quad \text{et} \quad \mathbb{V}(X) = np(1-p)$$

Exemples numériques





Distribution Multinomiale

- Écriture sous la forme $\mathcal{M}(n, p_1, \ldots, p_m)$, avec $n \in \mathbb{N}$ et $p_i \in [0, 1]$ $(i = \{1, \ldots, m\})$ les paramètres de la loi.
- Loi de probabilité :

$$\mathbb{P}(X_1 = x_1, \ldots, X_m = x_m) = n! \prod_{i=1}^{m} \frac{p_i^{x_i}}{x_i!}$$

avec $\Omega = [0, n]$, sachant que $\sum_i x_i = n$ et $\sum_i p_i = 1$.

- Moyenne et variance de chaque v.a. :

$$\mathbb{E}(X_i) = np_i \quad \text{et} \quad \mathbb{V}(X_i) = np_i(1 - p_i)$$

Distribution de Poisson

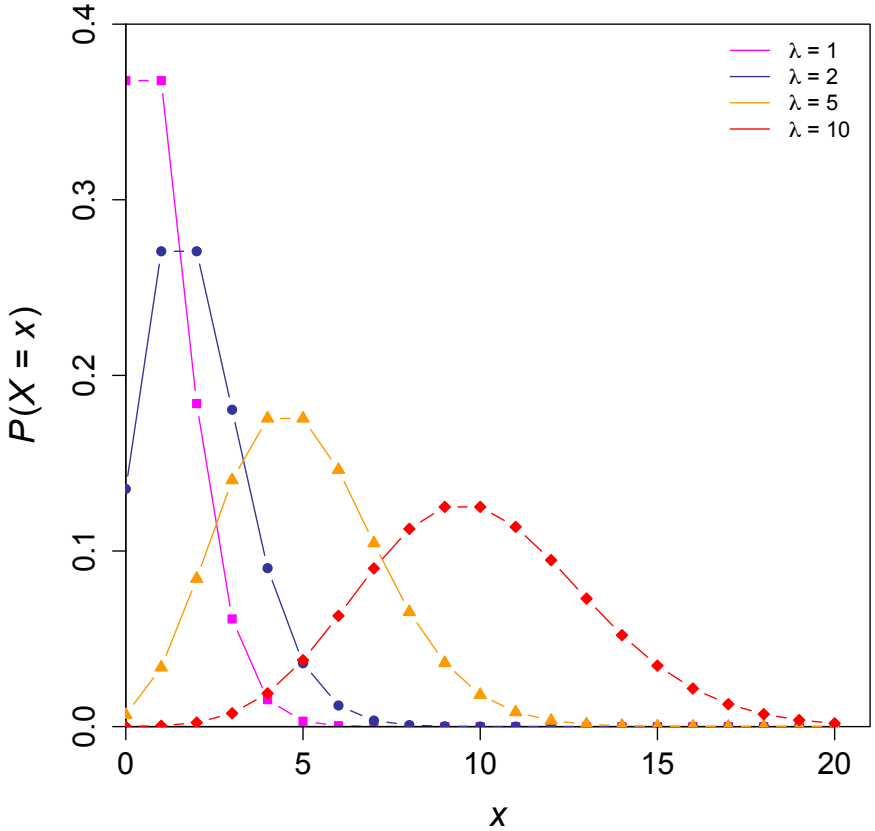
- Écriture sous la forme $\mathcal{P}(\lambda)$, avec $\lambda \in \mathbb{R}^{+*}$ le paramètre de la loi.
- Loi de probabilité :

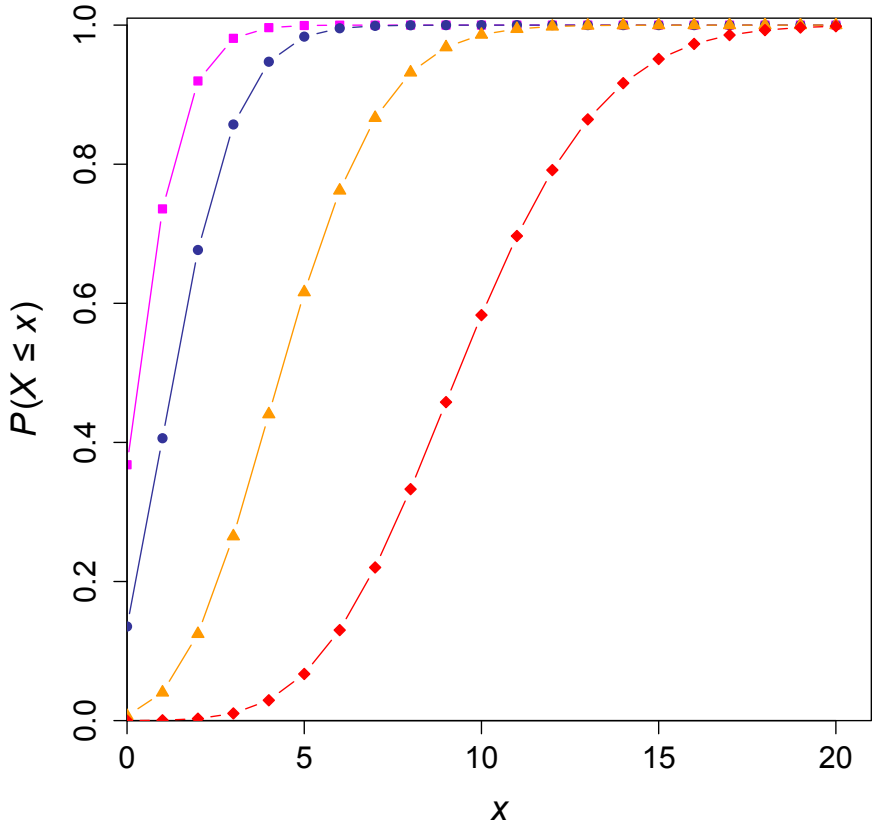
$$\mathbb{P}(X = x) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

  avec $\Omega = \mathbb{N}$.
- Moyenne et variance :

$$\mathbb{E}(X) = \mathbb{V}(X) = \lambda$$

Exemples numériques





Variables aléatoires continues

- Les valeurs prises par les variables appartiennent à des ensembles continus :
  - La probabilité d'un point est nulle.
  - Raisonnement en termes d'intervalles :

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x \quad a, b \in \Omega^2$$

- La loi d'une variable continue est définie par sa densité de probabilité f ou par sa fonction de répartition F :

$$\int_\Omega f(x)\mathrm{d}x = 1 \text{ et } F(x) = \mathbb{P}(X \leq x)$$

- La moyenne et la variance sont définies par :

$$\mathbb{E}(X) = \int_\Omega x\, f(x)\mathrm{d}x \text{ et } \mathbb{V}(X) = \int_\Omega [x - \mathbb{E}(X)]^2 f(x)\mathrm{d}x$$

Distribution Normale

- Écriture sous la forme $\mathcal{N}(\mu, \sigma^2)$, avec $\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^{+*}$ les deux paramètres de la loi.
- Loi de probabilité :

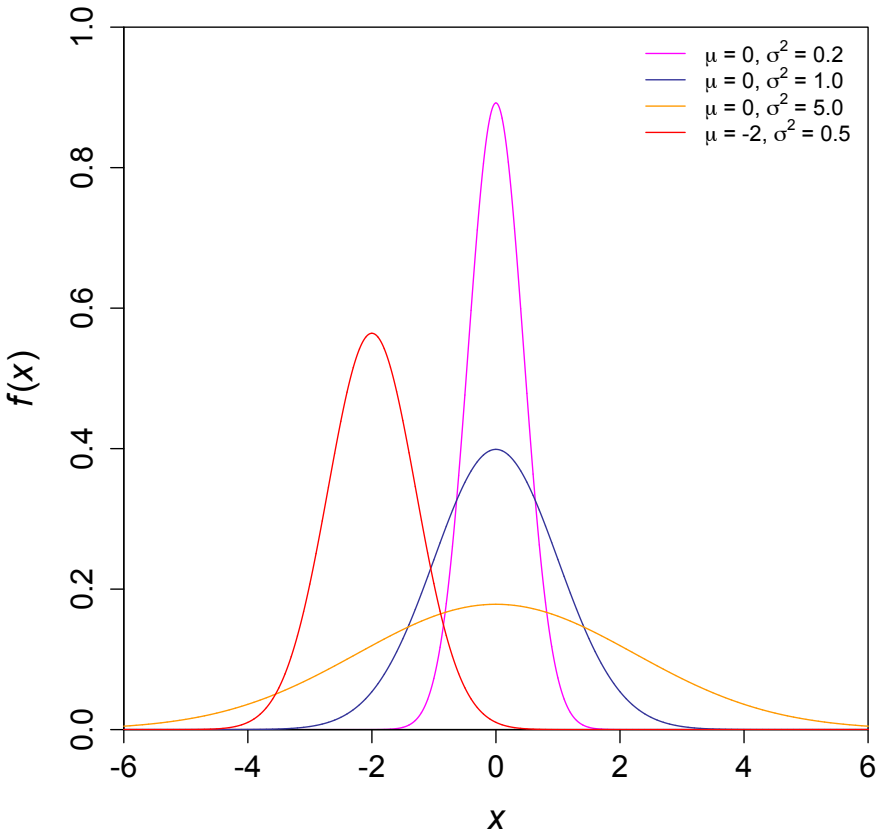
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

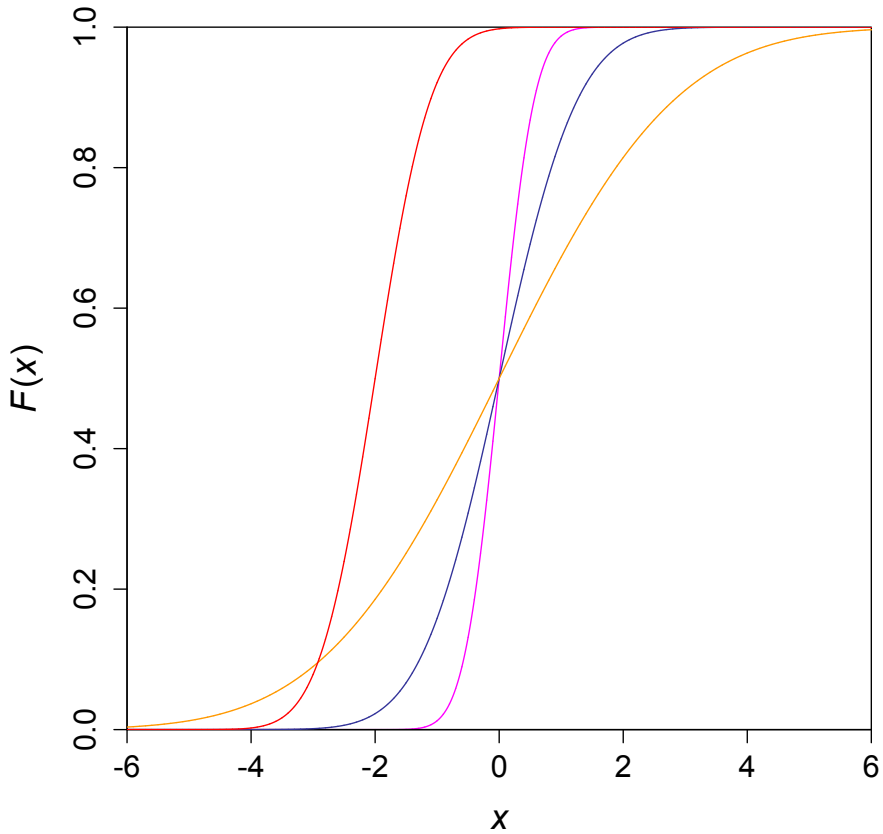
avec $\Omega = \mathbb{R}$.

- Moyenne et variance :

$$\mathbb{E}(X) = \mu \quad \text{et} \quad \mathbb{V}(X) = \sigma^2$$

Exemples numériques





Distribution Exponentielle

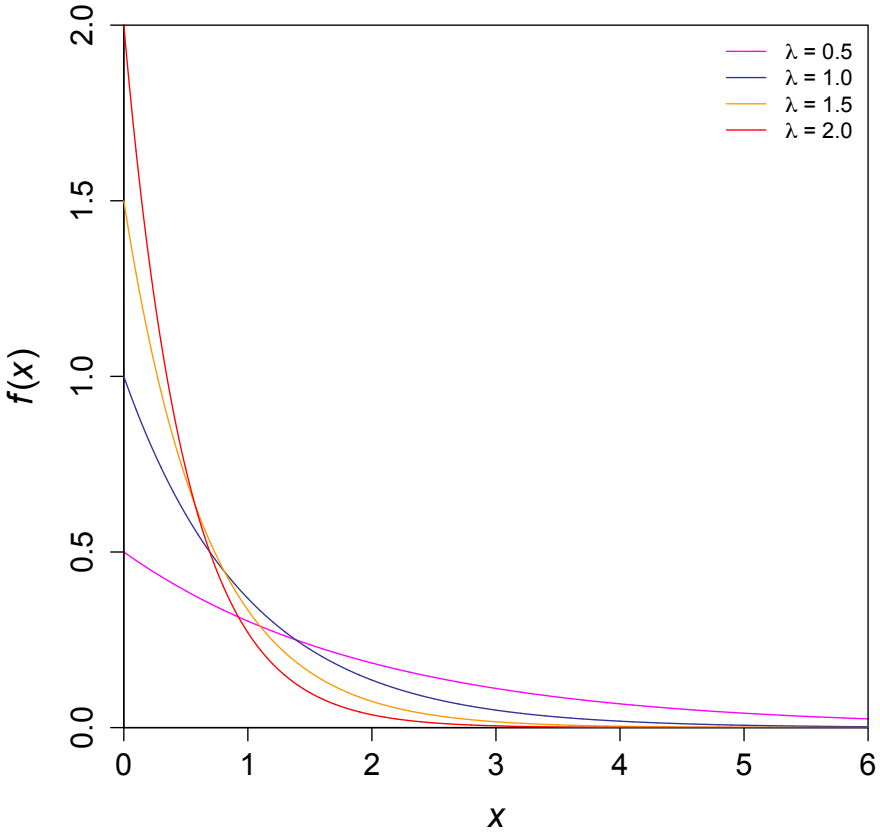
- Écriture sous la forme $\mathcal{E}(\lambda)$, avec $\lambda \in \mathbb{R}^{+*}$ le paramètre de la loi.
- Loi de probabilité :

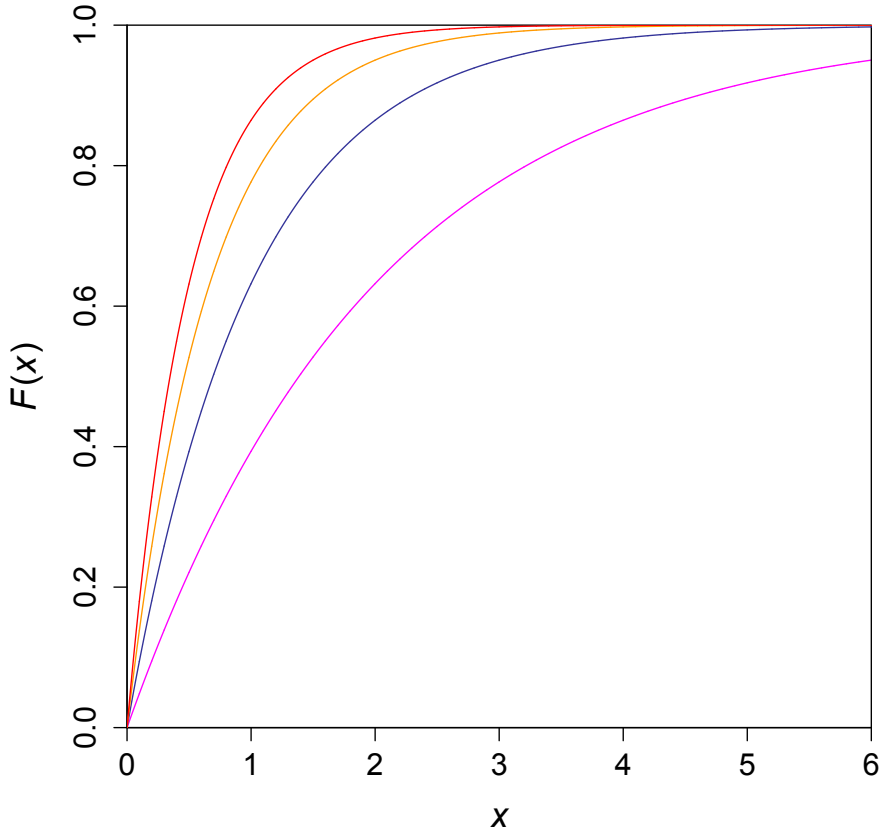
$$f(x) = \lambda \mathrm{e}^{-\lambda x}$$

  avec $\Omega = \mathbb{R}^{+}$.
- Moyenne et variance :

$$\mathbb{E}(X) = \frac{1}{\lambda} \quad \text{et} \quad \mathbb{V}(X) = \frac{1}{\lambda^2}$$

Exemples numériques





Distribution Gamma

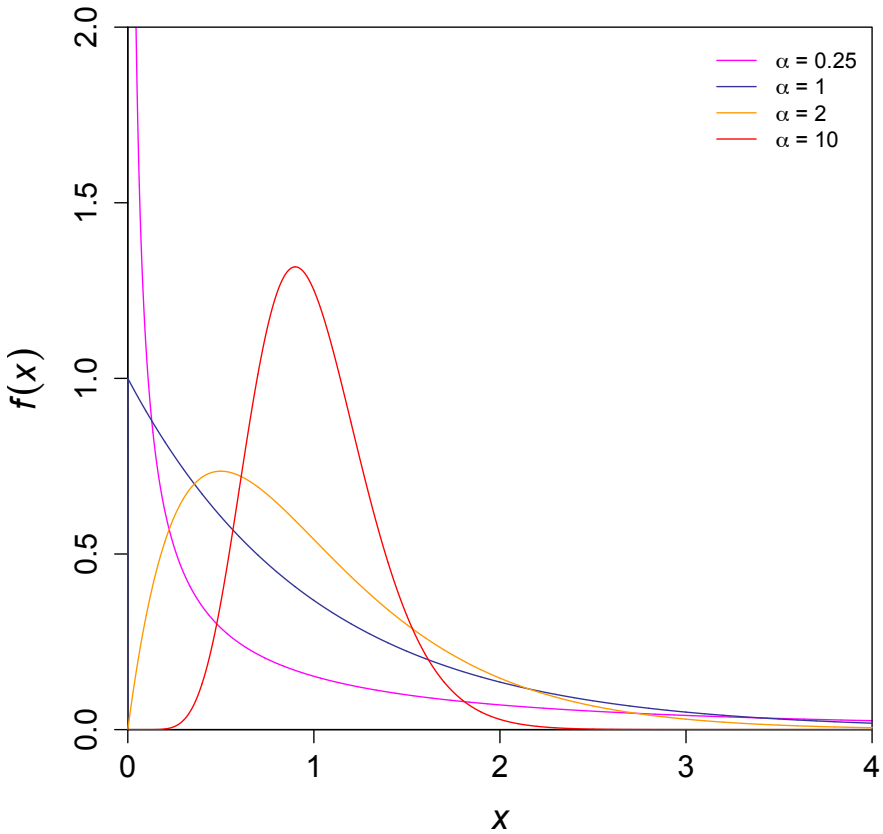
- Écriture sous la forme $\mathcal{G}(\alpha, \beta)$, avec $\alpha \in \mathbb{R}^{+*}$ et $\beta \in \mathbb{R}^{+*}$ les paramètres de la loi.
- Loi de probabilité :

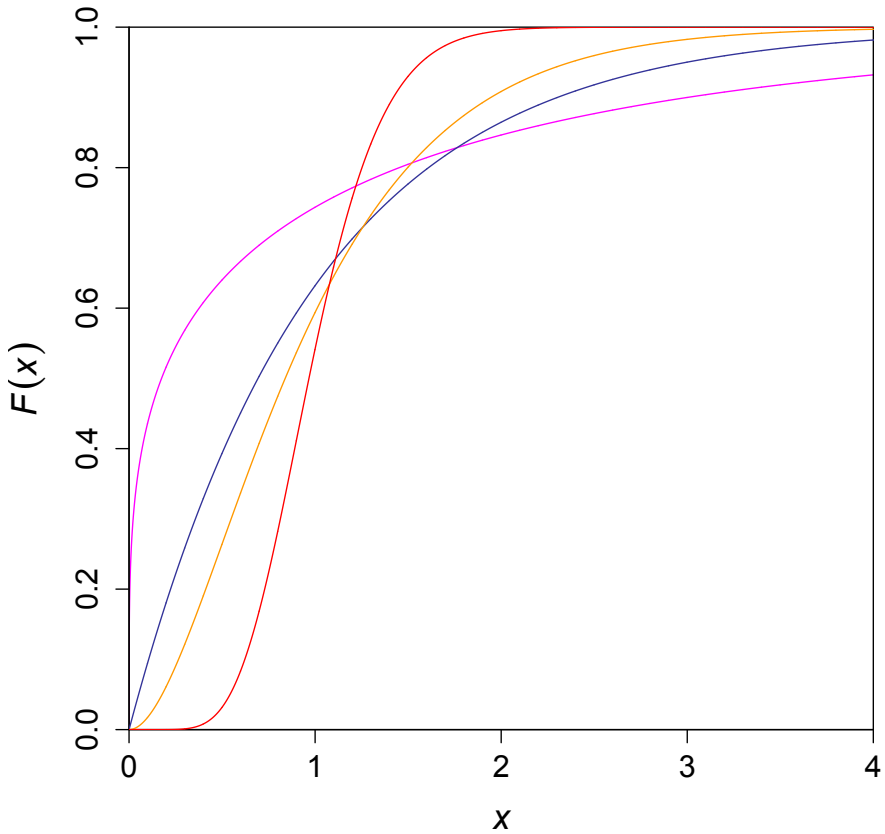
$$f(x) = \frac{x^{\alpha-1}\mathrm{e}^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \quad \text{avec} \quad \Gamma(\alpha) = \int_0^{\infty} \mathrm{e}^{-t} t^{\alpha-1} \mathrm{d}t$$

  avec $\Omega = \mathbb{R}^{+}$.
- Moyenne et variance :

$$\mathbb{E}(X) = \alpha\beta \quad \text{et} \quad \mathbb{V}(X) = \alpha\beta^2$$

Exemples numériques





Distribution du χ^2

- Écriture sous la forme $\chi^2(k)$ ou χ^2_k, avec $k \in \mathbb{N}^*$ le nombre de d.d.l. de la distribution.
- Loi de probabilité :

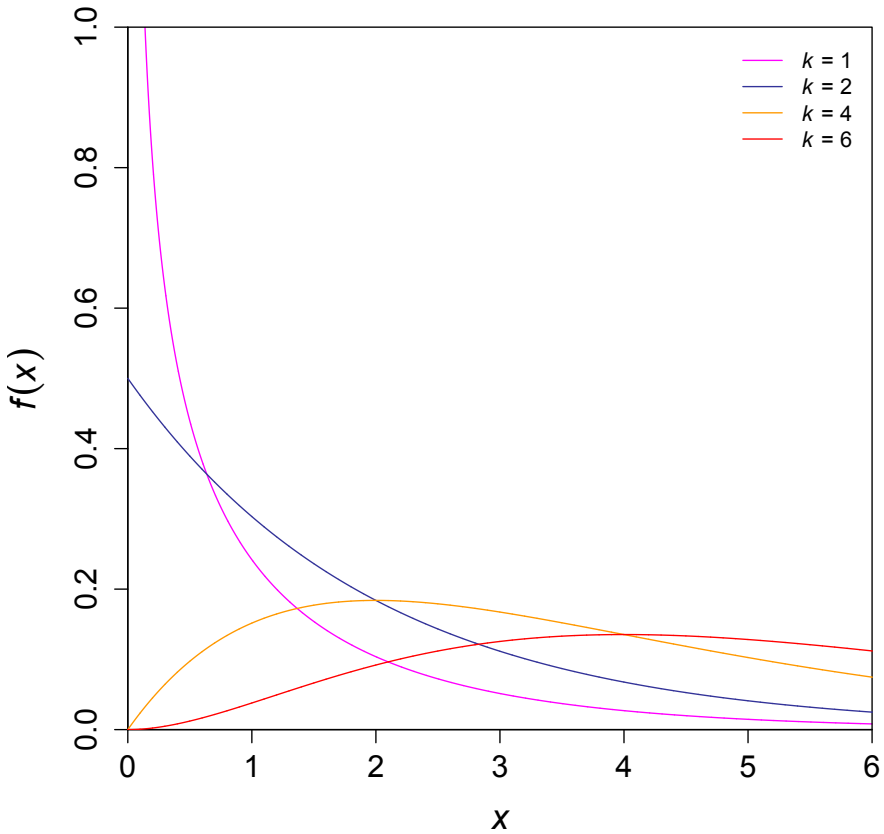
$$f(x) = \frac{x^{k/2-1}\mathrm{e}^{-x/2}}{2^{k/2}\Gamma(k/2)}$$

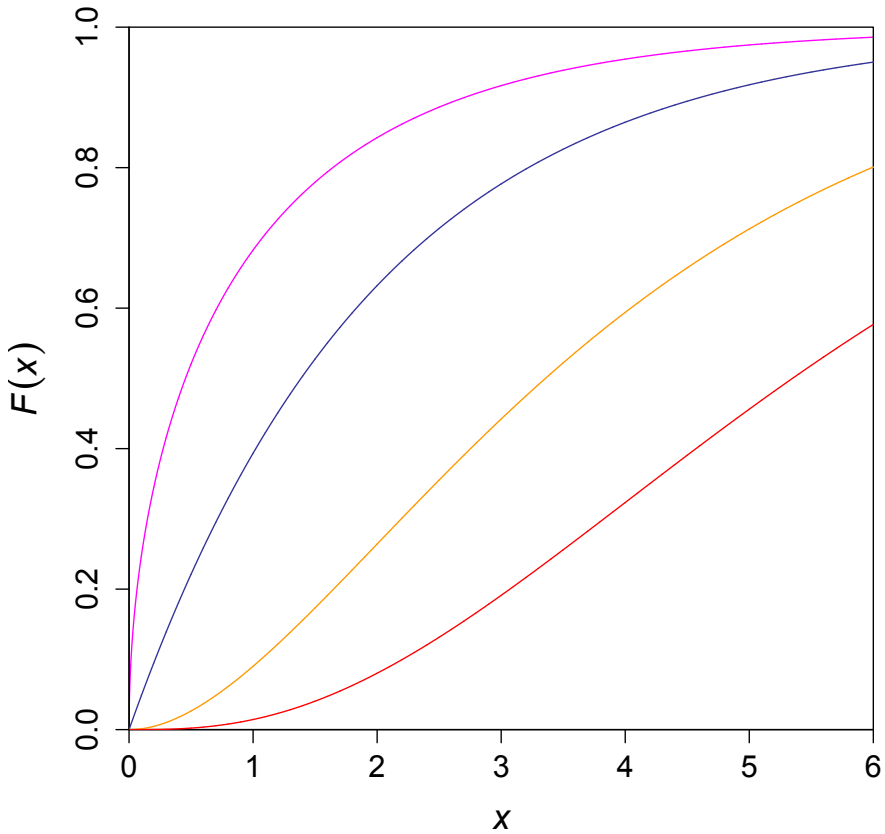
avec $\Omega = \mathbb{R}^+$.

- Moyenne et variance :

$$\mathbb{E}(X) = k \quad \text{et} \quad \mathbb{V}(X) = 2k$$

Exemples numériques





Distribution Uniforme

- Écriture sous la forme $\mathcal{U}(a, b)$, avec $a, b \in \mathbb{R}^2$ $(a < b)$ les paramètres de la loi.
- Loi de probabilité :

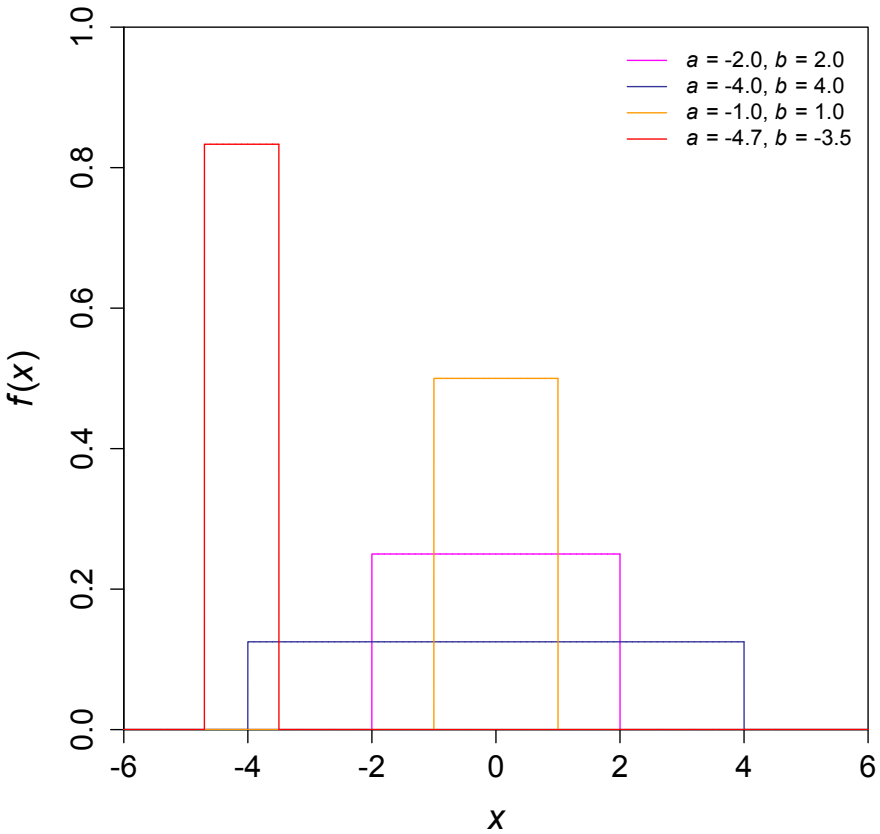
$$f(x) = \frac{1}{b-a} \; \forall x \in [a, b]$$

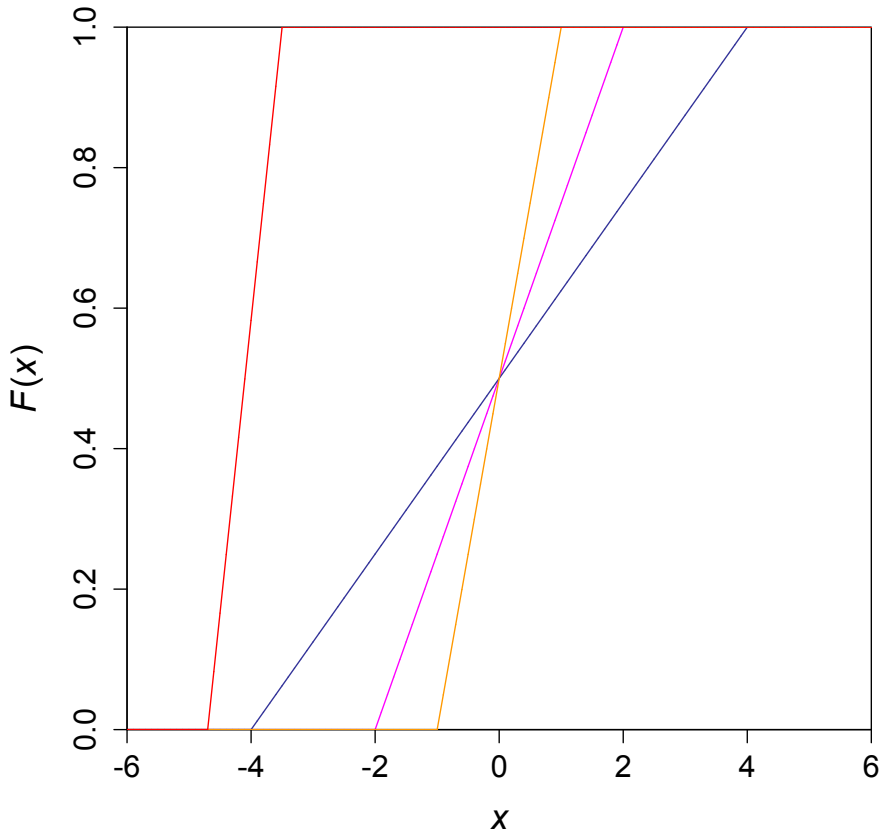
avec $\Omega = [a, b]$.

- Moyenne et variance :

$$\mathbb{E}(X) = \frac{a+b}{2} \quad \text{et} \quad \mathbb{V}(X) = \frac{(b-a)^2}{12}$$

Exemples numériques





Lectures conseillées

- En français :
  - Perrière G. et Brochier-Armanet C. (2010) *Concepts et Méthodes en Phylogénie Moléculaire*. Springer-Verlag, Paris.
- En anglais :
  - Felsenstein J. (2002) *Inferring Phylogenies*. Sinauer Associates, Sunderland.
  - Graur D. et Li W.H. (2000) *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland.
  - Nei M. et Kumar S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
  - Yang Z. (2006) *Computational Molecular Evolution*. Oxford University Press, New York.