Evaluation et tests des phylogénies

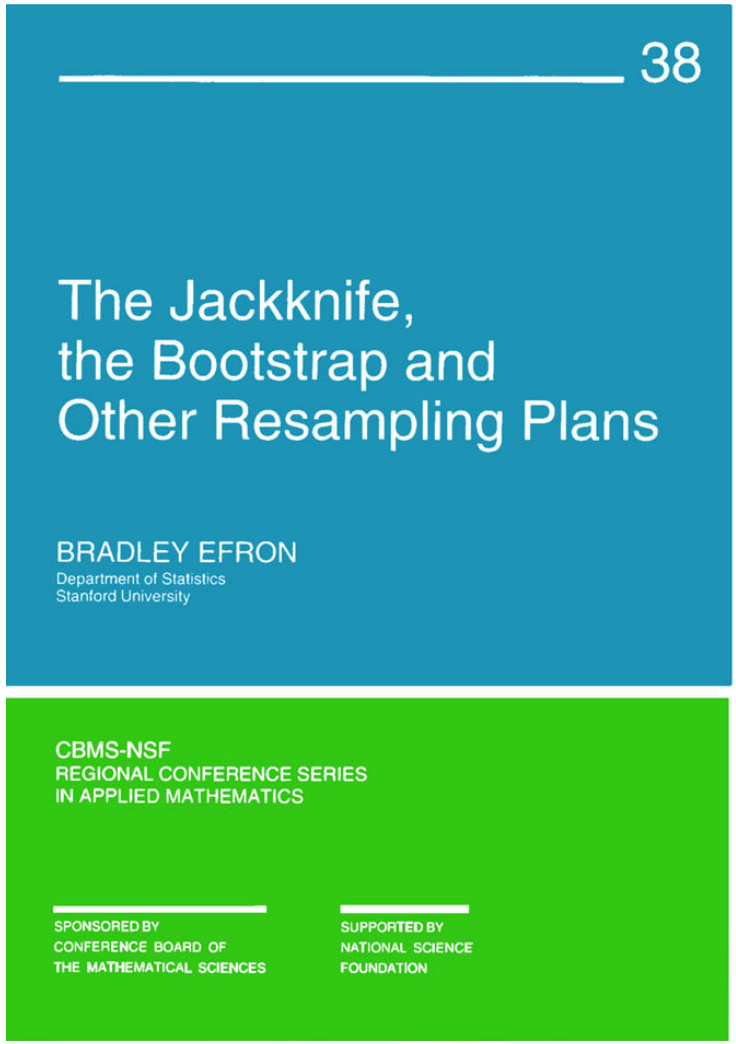
Guy Perrière & Héloïse Philippon

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

31 novembre 2017

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.

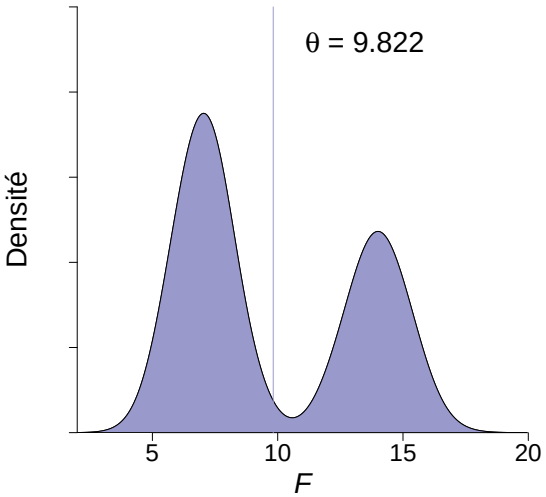


Principe général

- 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.

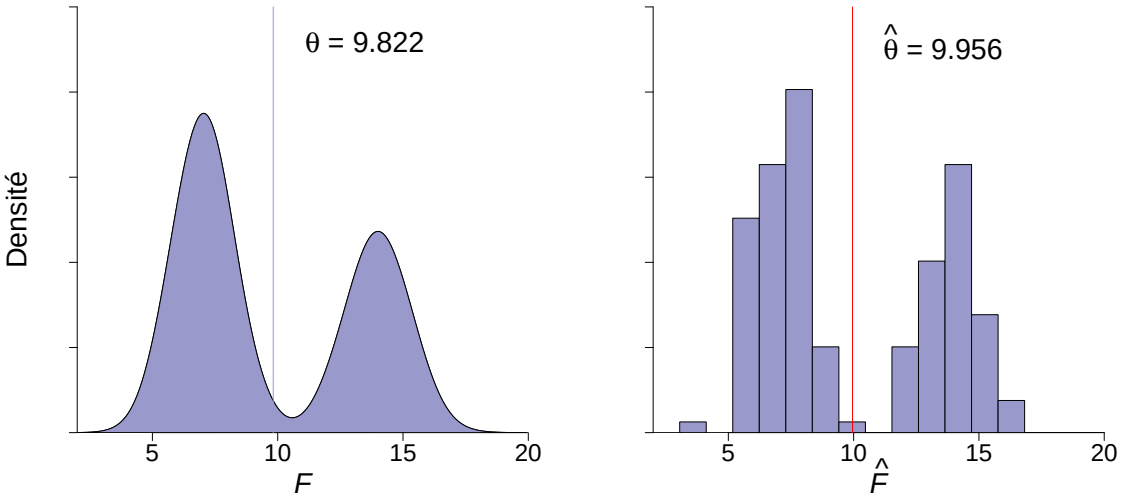
Moyenne d'une distribution

- 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 : $\theta = 9.822$.



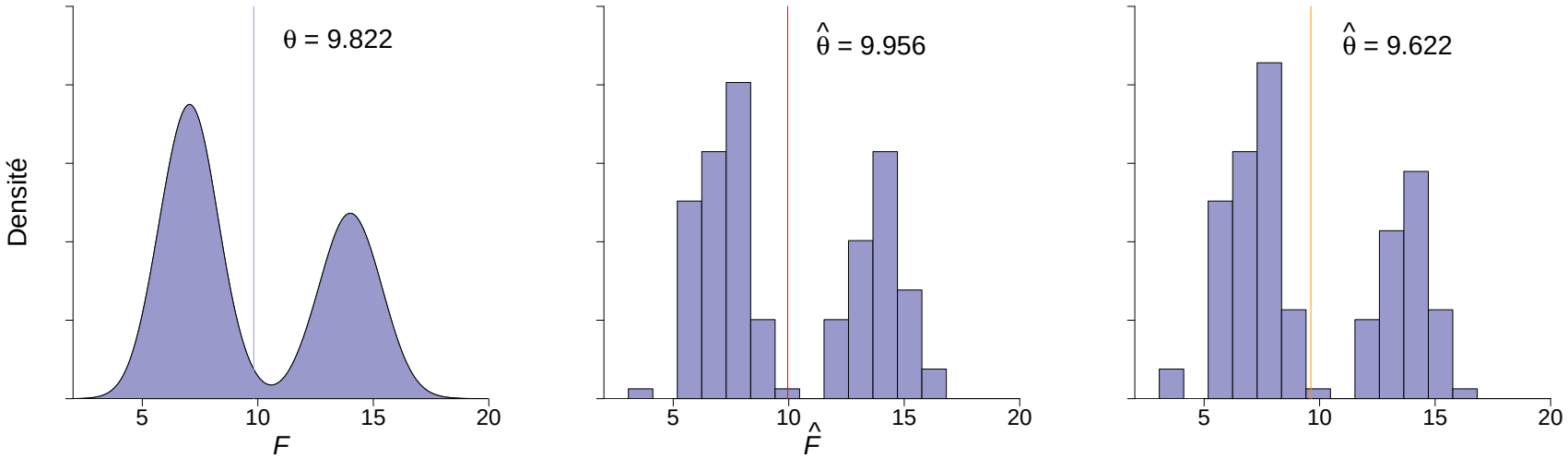
Moyenne d'une distribution

- 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 : $\hat{\theta} = 9.956$.



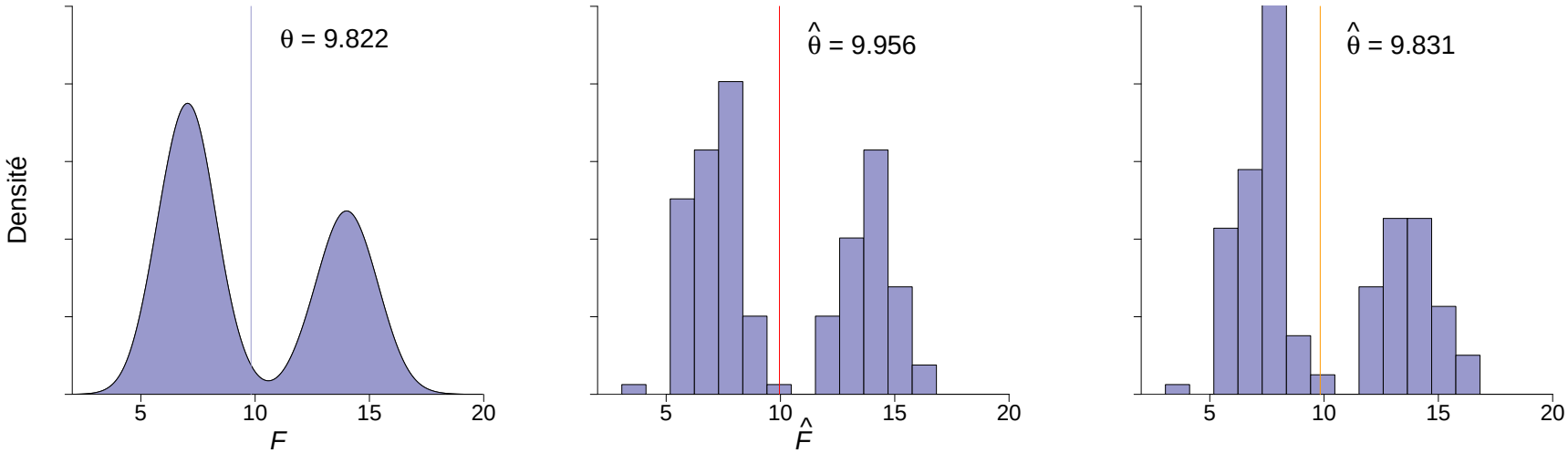
Moyenne d'une distribution

- 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* :



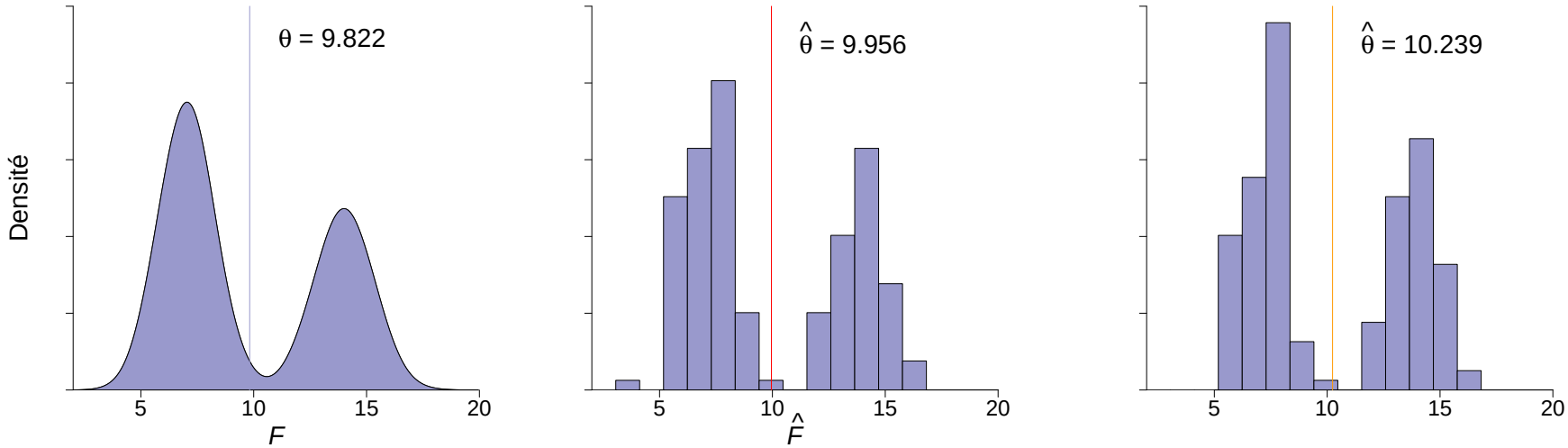
Moyenne d'une distribution

- 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* :



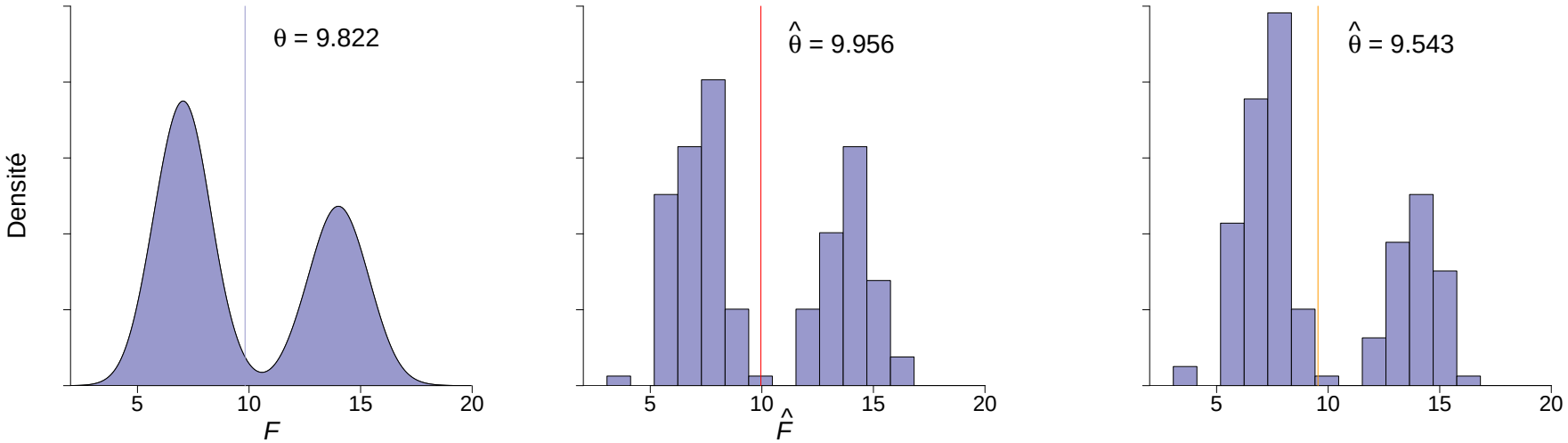
Moyenne d'une distribution

- 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* :

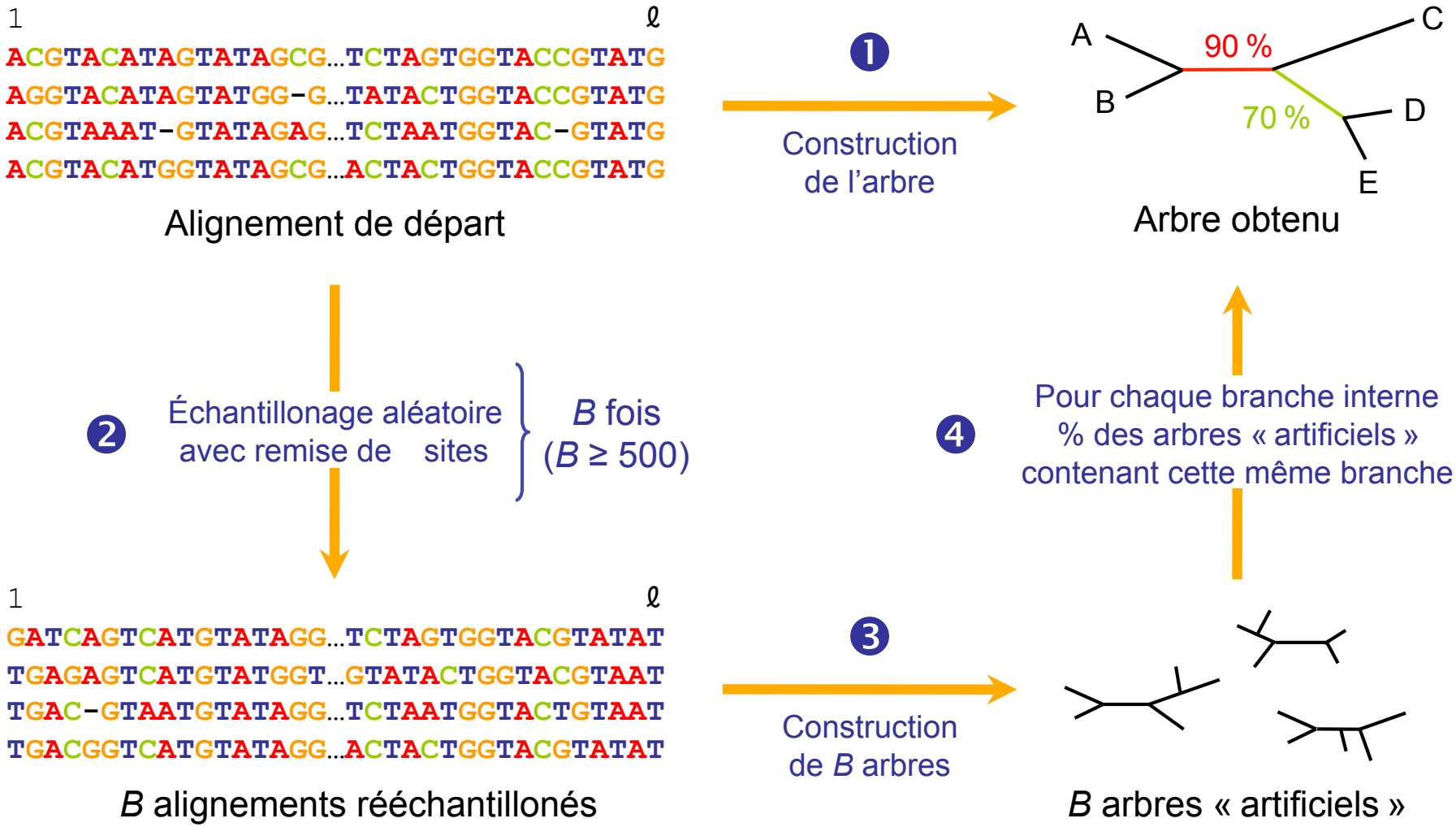


Moyenne d'une distribution

- 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.

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 $(\mathbf{b}, \boldsymbol{\vartheta}, \alpha)$.
  - 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 :

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

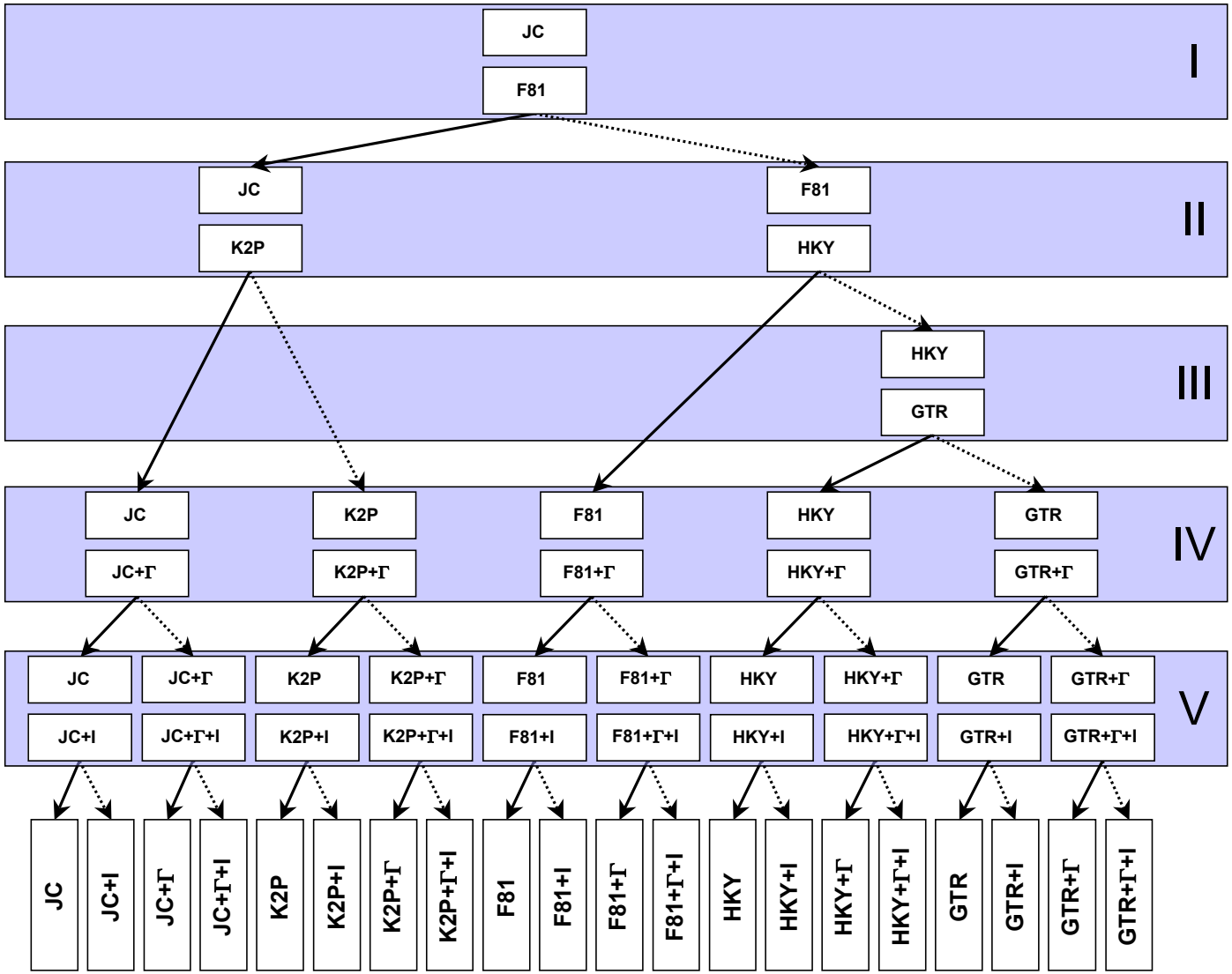
Likelihood Ratio Test (LRT)

- 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 :

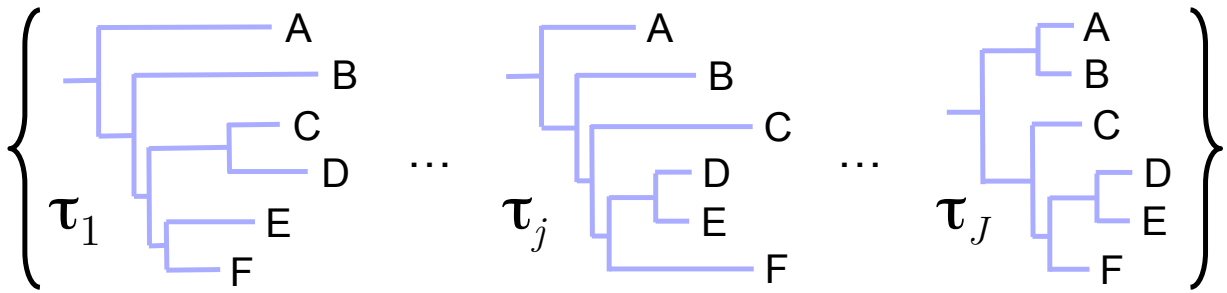
$$\text{BIC} = -2 \ln L(\boldsymbol{\vartheta}) + k \ln \ell$$

- Comme dans le cas de l'AIC, sélection du modèle présentant la plus faible valeur au test.
- Approximation du test de comparaison de modèles utilisant les Facteurs de Bayes (*cf.* cours sur l'inférence bayésienne) :

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

Nécessité

- 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 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 :

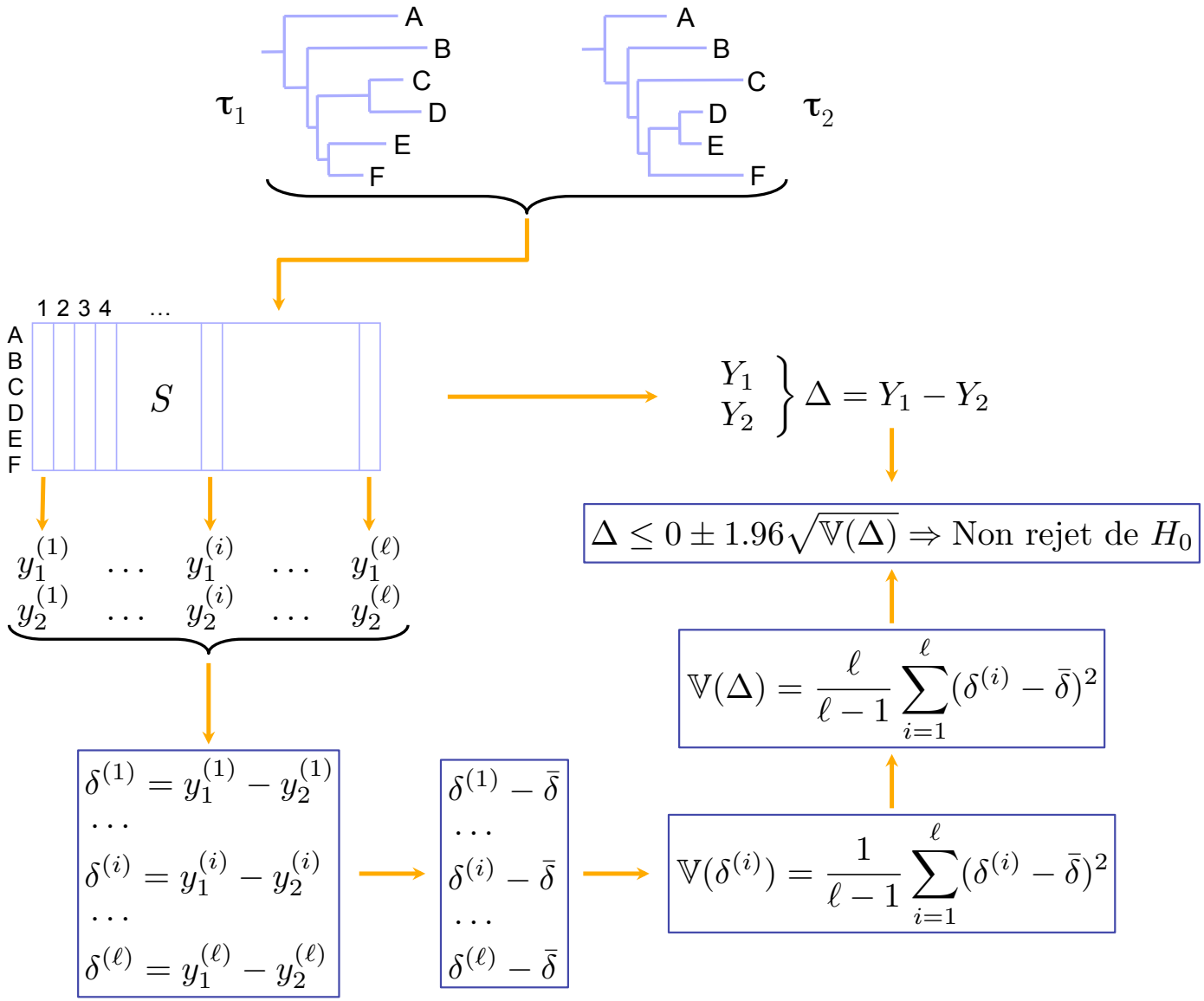
$$\bar{\delta} = \frac{1}{\ell} \sum_{i=1}^{\ell} \delta^{(i)} = \frac{\Delta}{\ell}$$

- 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 rééchantillonages 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 :

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

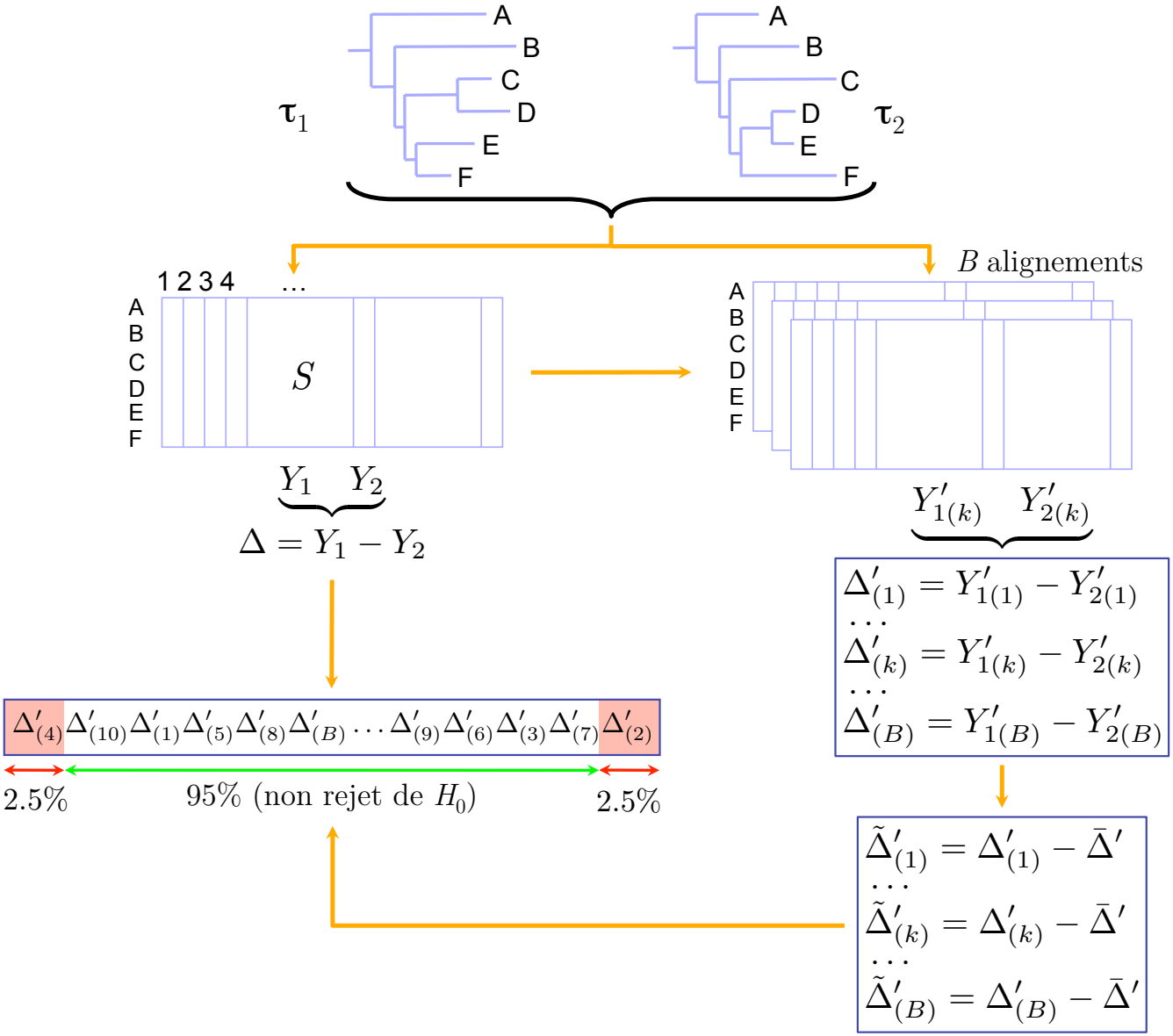
Approche par *bootstrap* (II)

- 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

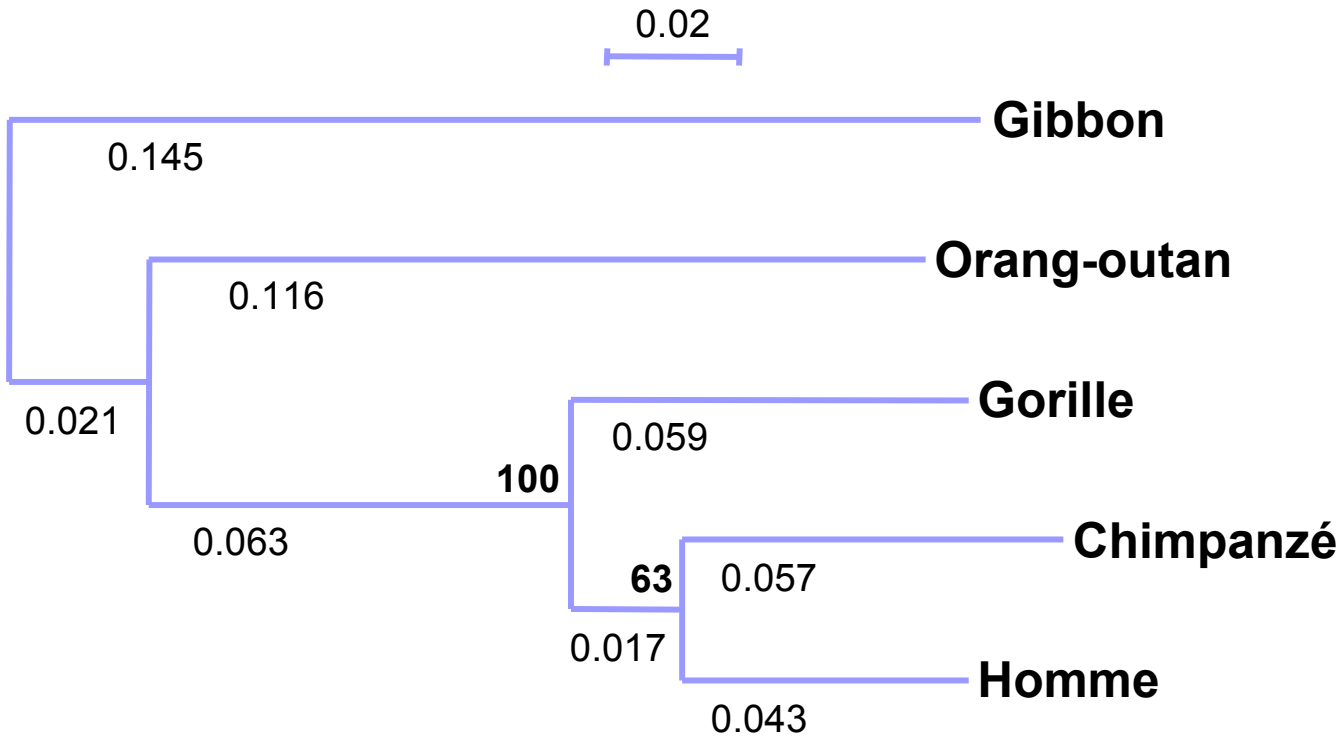


Limitations

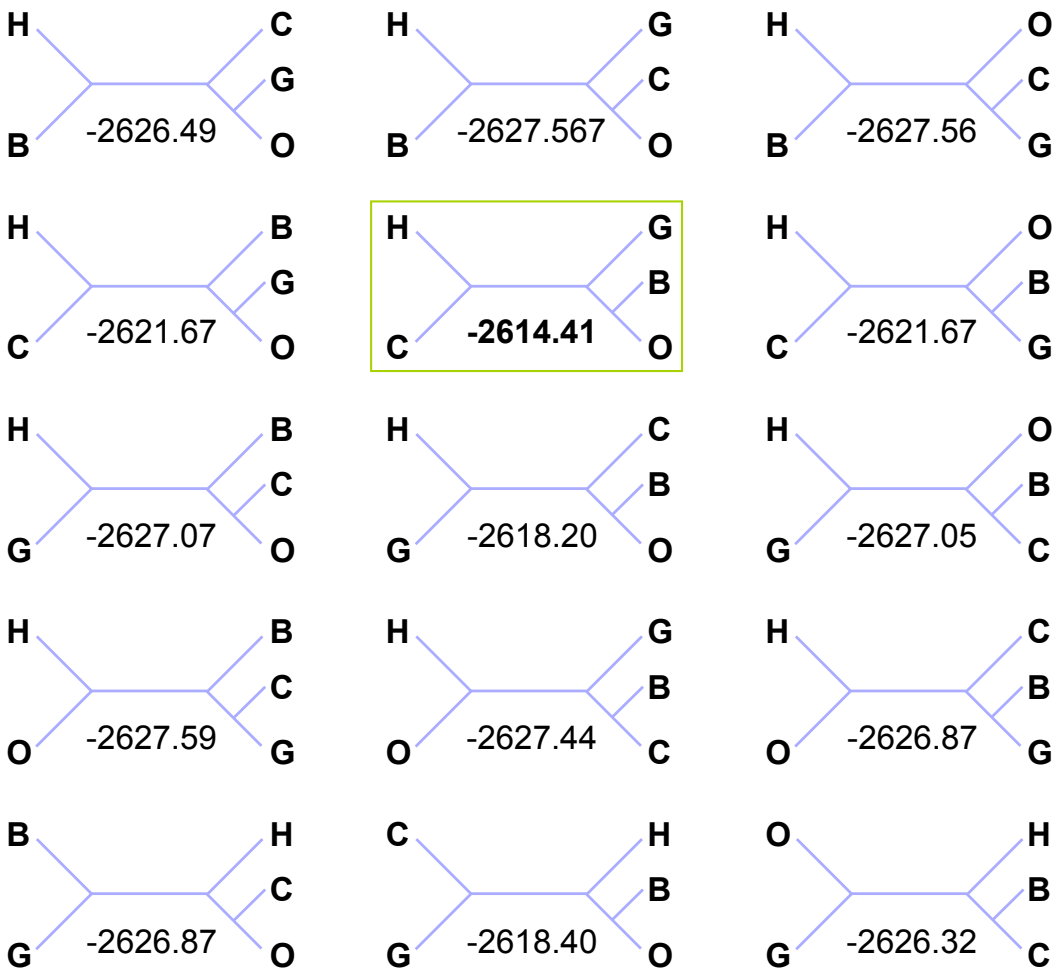
- Test limité à la comparaison de deux topologies :
  - Pas de correction pour les tests multiples.
- Les arbres testés doivent êtres choisis *indépendamment* des données utilisées pour réaliser le test :
  - Indispensable pour justifier l'hypothèse nulle sous laquelle $\mathbb{E}(\Delta) = 0$.
  - Le 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

- 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

Comparaison des topologies

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