Délimitation d'espèces potentielles grâce à des données génétiques pour un locus

Marc Manceau

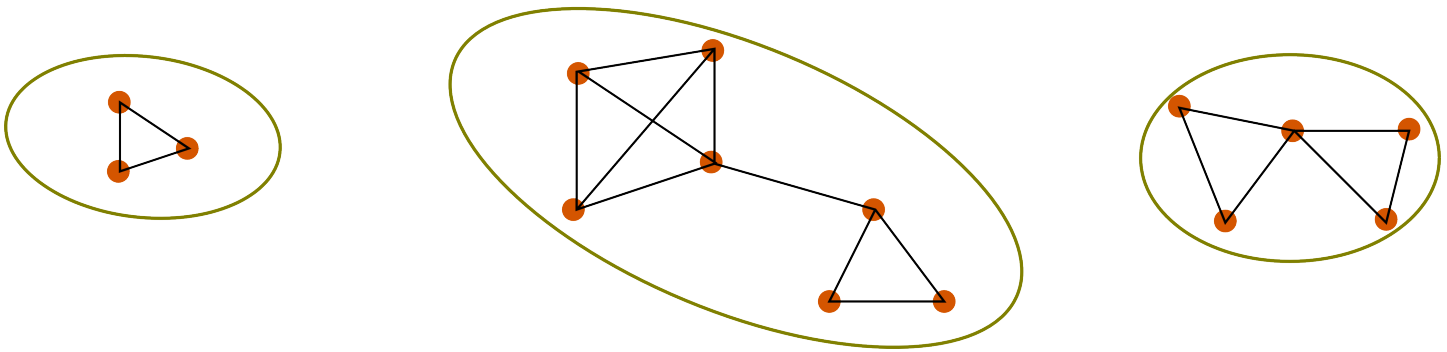
Journal club *SMILE*
Stochastic Modeling for the Inference of Life Evolution

5 Mai 2015

Motivation

- On est toujours autant focalisé sur le rang d'espèce.
- Beaucoup d'espèces difficilement distinguables morphologiquement.
- Gain de temps pour identification et description.

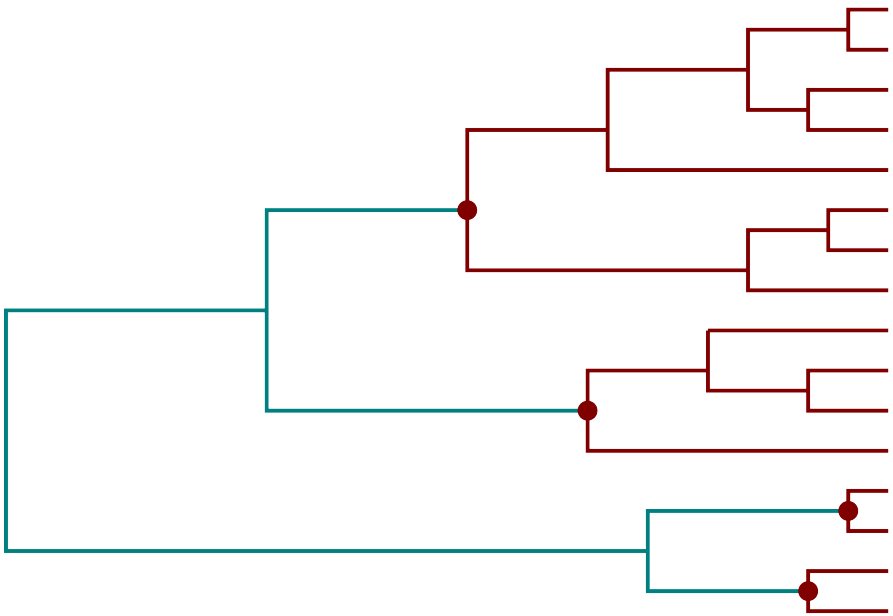
Alignements 2 à 2			différences 2 à 2	Threshold : 2 différences
ATTGGGTGTA	vs	ATTGCGTGAA	2	
ATTGGGTGAA	vs	ATAGCGTGGA	3	
TTTGCGTGAA	vs	AATGGGTTAG	5	
⋮		⋮	⋮	
ATTACGTGAA	vs	ATTGCGTGAA	1	



Principe de délimitation un peu automatisé

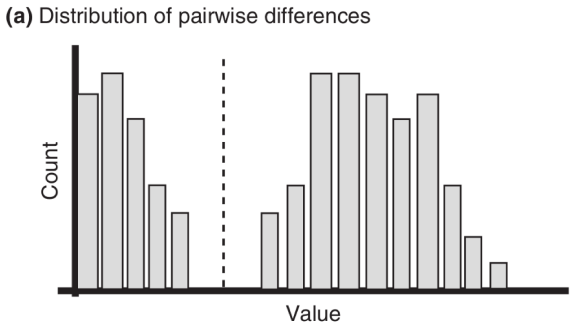
- Le moins possible d'input de la part de l'utilisateur.
- Détecter la signature du passage d'une divergence "intra-population" à une divergence "inter-espèce" si elle existe.

Délimitation sur arbre reconstruit



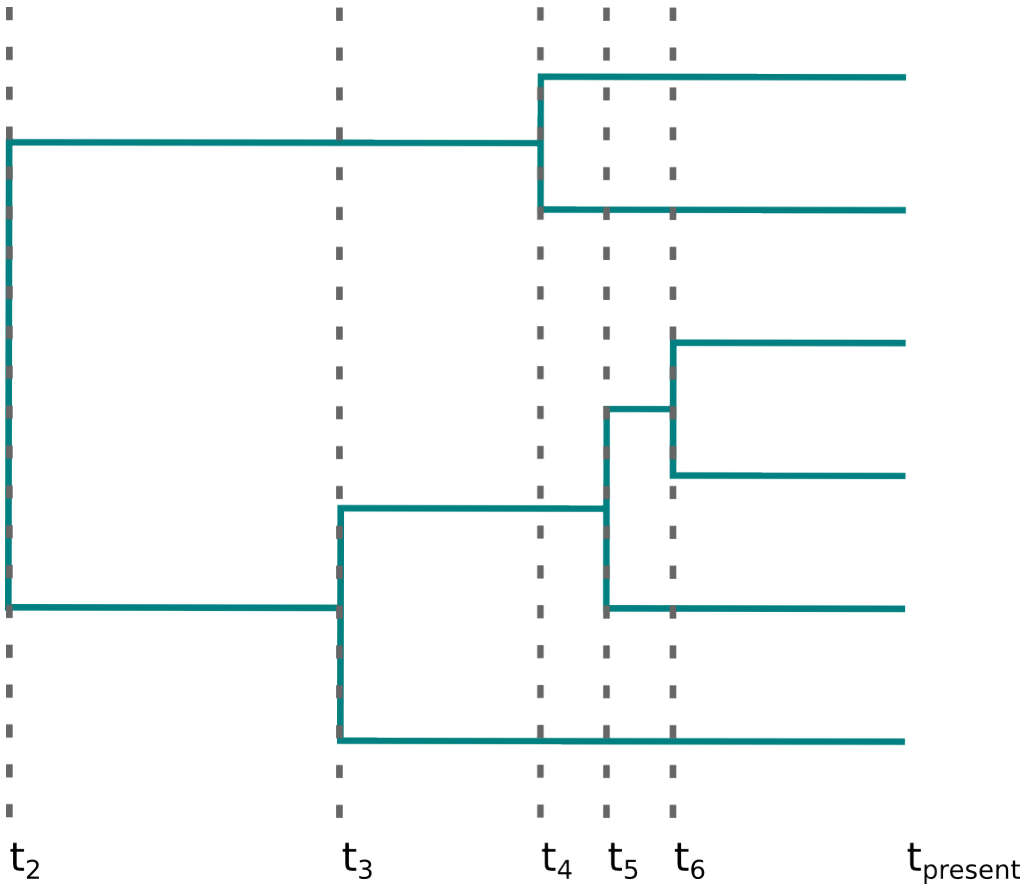
Délimitation sur alignement de séquences

ATTGGGTGTA vs ATTGCGTGAA
ATTGGGTGAA vs ATAGCGTGGA
TTTGCGTGAA vs AATGGGTTAG
⋮ ⋮
ATTACGTGAA vs ATTGCGTGAA



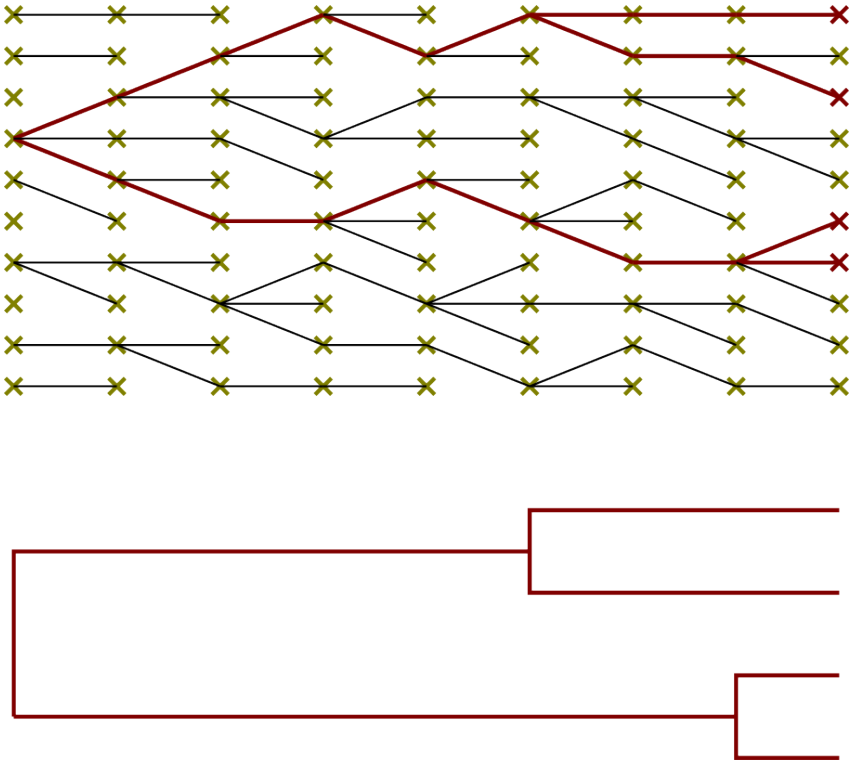
Arbre des espèces : Yule

- Chaque lignée donne naissance après un temps $\sim \mathcal{E}(\lambda_{\text{spec}})$.
- Avec $n_{i,\text{spec}}$ lignées, le prochain temps de coalescence $\sim \mathcal{E}(\lambda_{\text{spec}} n_{i,\text{spec}})$.
- Les deux branches qui coalescent sont choisies uniformément.



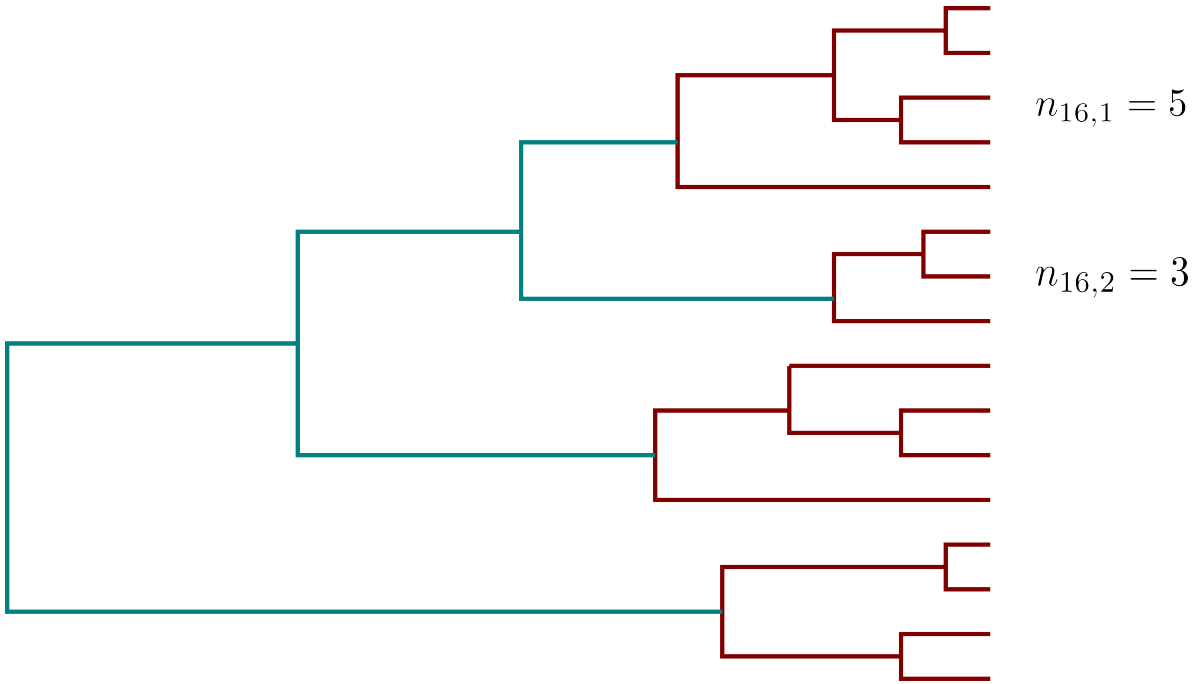
Coalescences intra-espèce : Kingman

- Modèle classique de coalescence en population panmictique.
- Avec $n_{i,j}$ lignées, le prochain temps de coalescence $\sim \mathcal{E}(\lambda_j n_{i,j}(n_{i,j}-1))$.
- Les deux branches qui coalescent sont choisies uniformément.



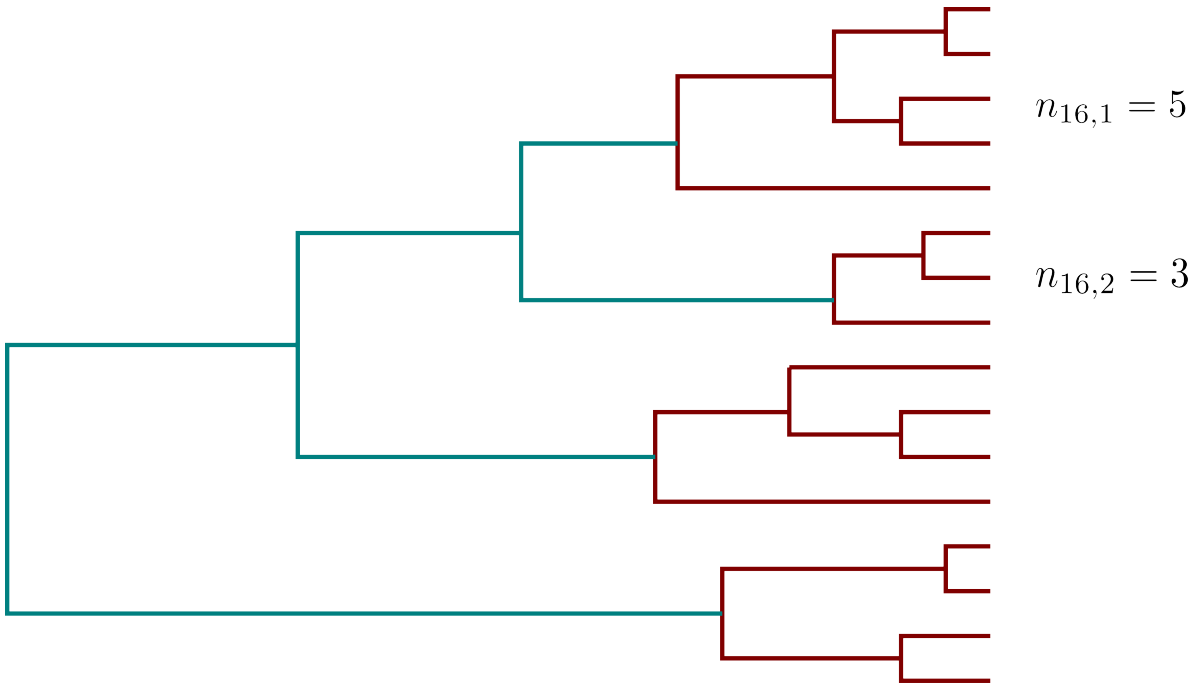
Deux modèles ensemble

- Squelette interne de Yule, Kingman indépendants aux extrémités.
- Il n'y a pas de taux de passage de l'un à l'autre.
- Avec $(n_{i,\text{spec}}, n_{i,1}, \ldots n_{i,k})$ lignées, le prochain temps de coalescence $\sim \mathcal{E}\left(\lambda_{\text{spec}} n_{i,\text{spec}} + \sum_{j=1}^{k} \lambda_j n_{i,j}(n_{i,j}-1)\right)$.
- Tous les branchements ne sont plus équiprobables (?)



Un rajout : le paramètre "p"

- Ce ne sont finalement plus des arbres de Yule ou de Kingman.
- Taux de coalescence inter-espèces : $\lambda_{\text{spec}} n_{i,\text{spec}}^{p_{\text{spec}}}$.
- Taux de coalescence intra-espèce j : $\lambda_j \left(n_{i,j}(n_{i,j}-1)\right)^{p_j}$.



Vraisemblance d'un arbre et estimation des paramètres

- Ils considèrent un unique taux de coalescence :

$$b_i^* = \lambda_{\text{spec}} n_{i,\text{spec}}^{p_{\text{spec}}} + \sum_{j=1}^{k} \lambda_{\text{coal}} \left(n_{i,j}(n_{i,j}-1)\right)^{p_{\text{coal}}}$$

- La vraisemblance qu'ils donnent correspond à :

$$\mathcal{L}(\text{arbre}) = \prod_{i=1}^{N} L(x_i) = \prod_{i=1}^{N} b_i^* e^{-b_i^* x_i} \text{ d'où } \ln \mathcal{L} = \sum_{i=1}^{N} \ln b_i^* - b_i^* x_i$$

- Si p_{spec} et p_{coal} sont connus, λ_{spec} et λ_{coal} sont estimés par :

$$\lambda_{\text{spec}} = \frac{\#\text{ branch. inter-espèces}}{\sum_{i=1}^{N} n_{i,\text{spec}}^{p_{\text{spec}}} x_i}$$

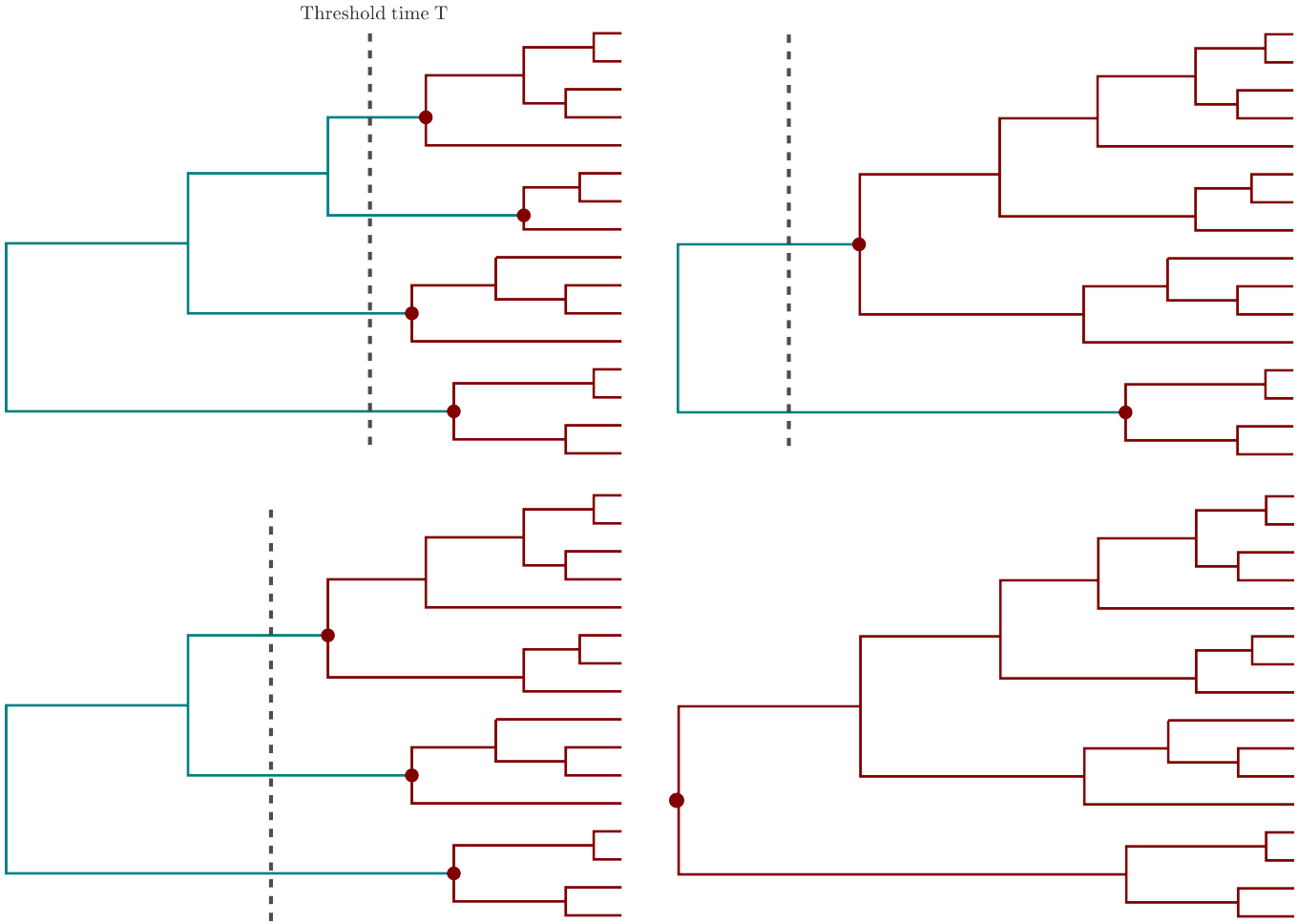
$$\lambda_{\text{coal}} = \frac{\#\text{ branch. intra-espèces}}{\sum_{i=1}^{N} \sum_{j=1}^{k} \left(n_{i,j}(n_{i,j}-1)\right)^{p_{\text{coal}}} x_i}$$

- p_{spec} et p_{coal} sont ensuite estimés numériquement par maximum de vraisemblance.

Optimisation de l'assignement des noeuds

"Single threshold"

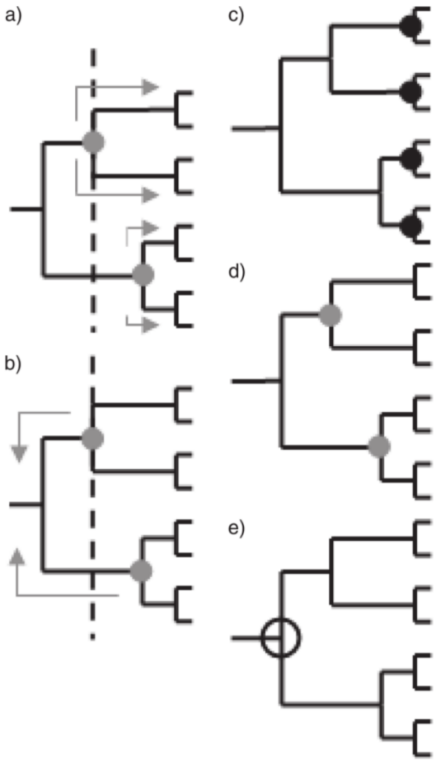
- Les espèces sont monophylétiques sur l'arbre de gène.
- Les événements de spéciation sont antérieurs aux coalescences intra-pop.

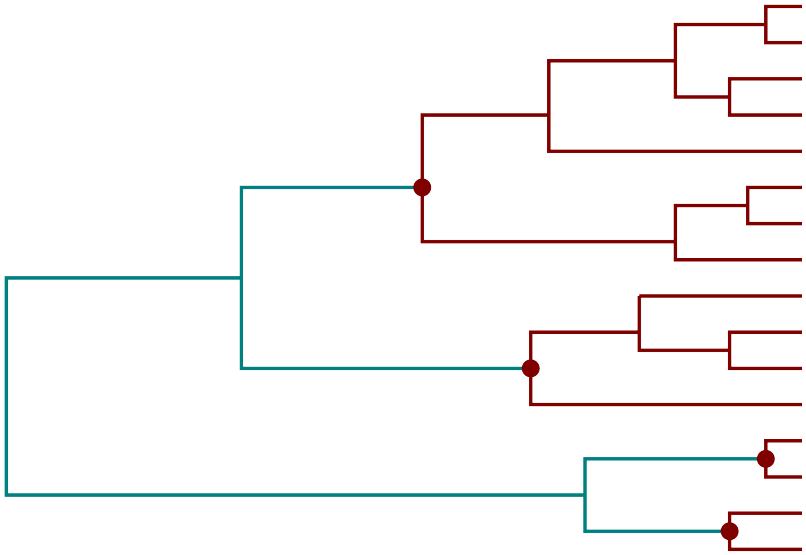


Optimisation de l'assignement des noeuds

"Multiple threshold"

- On commence avec plusieurs configurations aléatoires de MRCA, ou bien la meilleure configuration du "single threshold".
- Puis on teste les meilleures possibilités en divisant ou groupant des clusters.



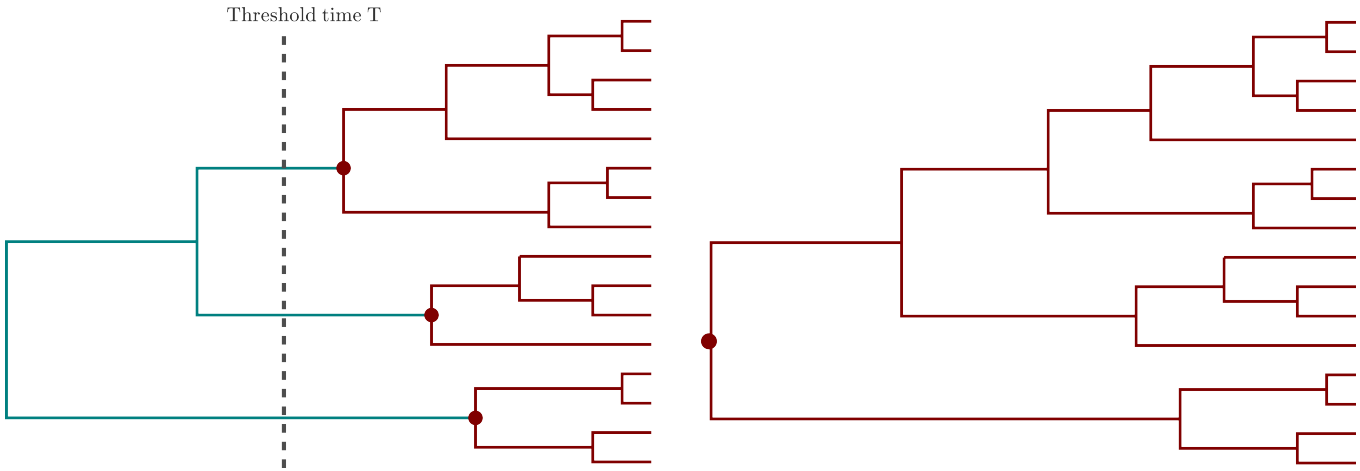


Test d'hypothèse

H0 tous les individus sont dans la même espèce.

H1 il y a au moins deux espèces différentes.

- Test de rapport de vraisemblance (sans détail).
- "The rate of false positives (rejecting the null hypothesis at a 95% level when it is true) were $\alpha = 0.02$ and $\alpha = 0.07$ for the single and multiple threshold methods".



Support associé à chaque noeud

- On propose un ensemble de modèles Θ, qui sont des arbres avec points de passage Yule-Kingman.
- Le score AIC_M et le poids d'Akaike w_M sont calculés pour chaque $M \in \Theta$:

$$\mathrm{AIC}_M = 2k - 2 \ln \max \mathcal{L}$$

$$w_M = \frac{e^{-\frac{1}{2}\left(AIC_M - \min\limits_{m\in\Theta} AIC_m\right)}}{\sum\limits_{n\in\Theta} e^{-\frac{1}{2}\left(AIC_n - \min\limits_{m\in\Theta} AIC_m\right)}}$$

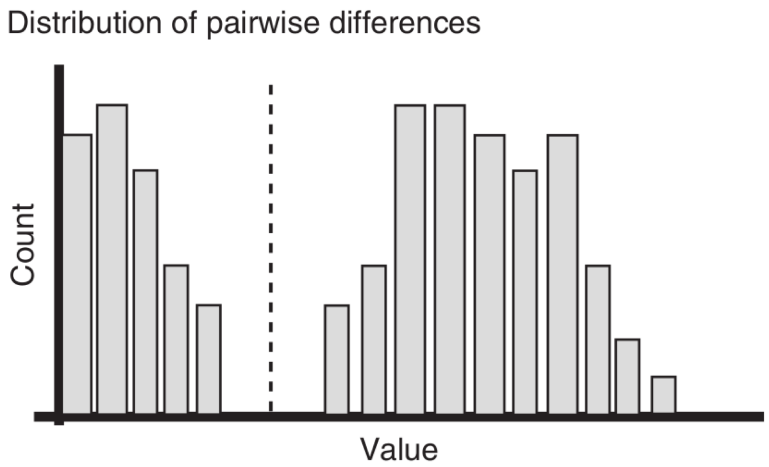
- Ils en déduisent un intervalle de confiance pour les modèles.
- Et un support pour qu'un noeud N_j soit MRCA :

$$w^{N_j}_{\mathrm{MRCA}} = \sum_{M\in\Theta} w_M \mathbf{1}_{N_j \text{ est un MRCA dans le modèle } M}$$

Le barcode gap

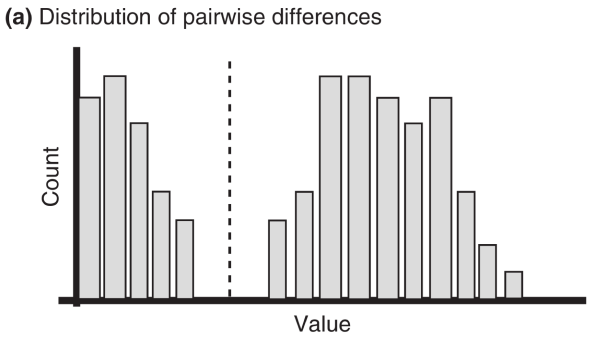
- Observons la distribution des différences deux à deux entre les barcodes.
- Un *barcode gap* est un trou dans cette distribution.

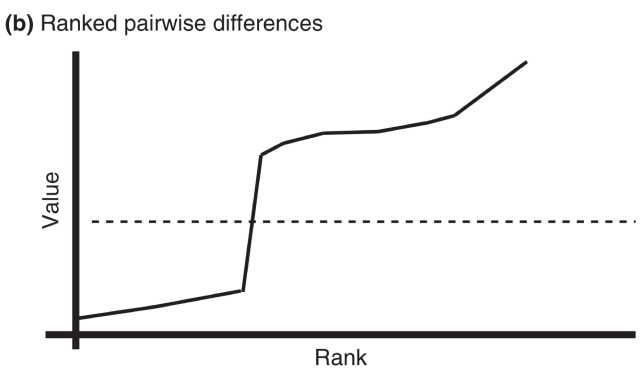
$\frac{n(n-1)}{2}$ comparaisons			différences 2 à 2
ATTGGGTGTA	vs	ATTGCGTGAA	2
ATTGGGTGAA	vs	ATAGCGTGGA	3
TTTGCGTGAA	vs	AATGGGTTAG	5
⋮		⋮	⋮
ATTACGTGAA	vs	ATTGCGTGAA	1

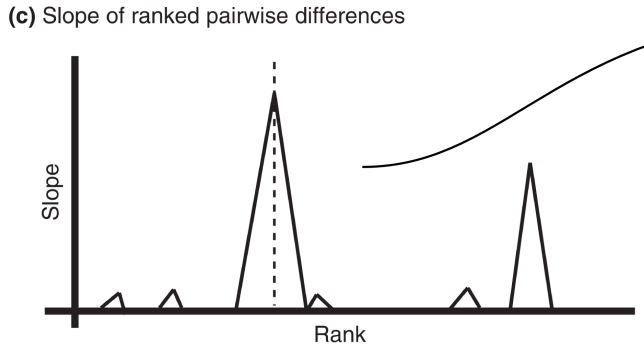


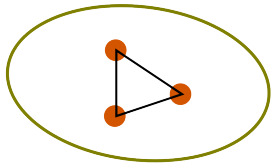
- Attendu en cas d'écart entre les distributions de différences 2 à 2 intra et inter-spécifiques.

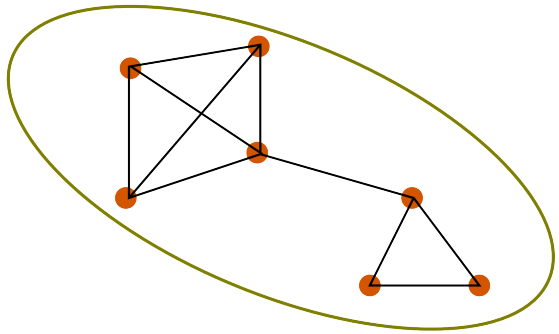
Principe général



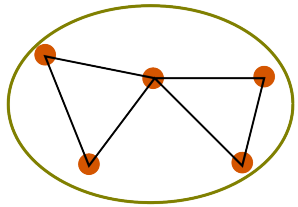




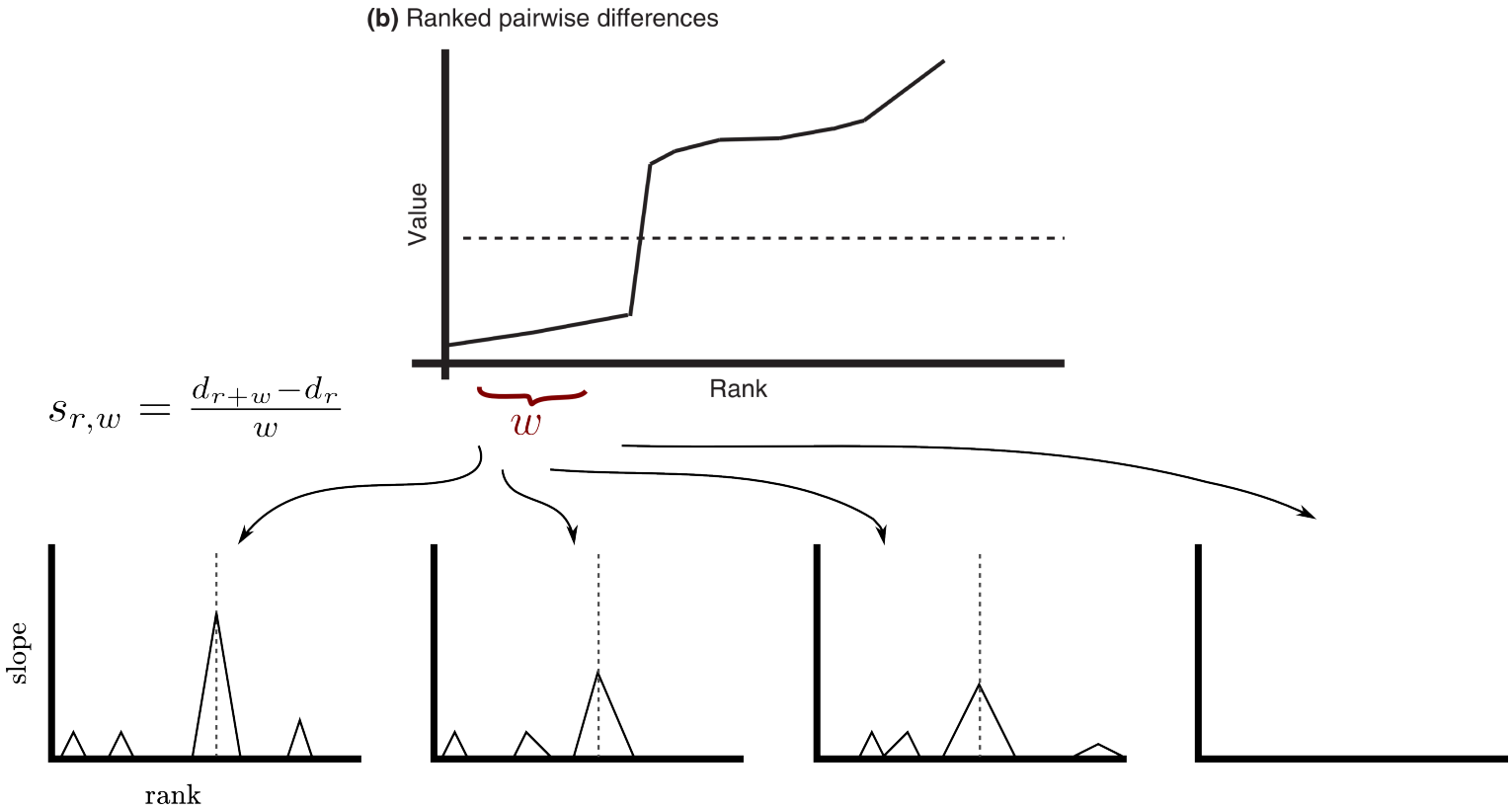




Deux séquences sont en relation si
leur différence est inférieure à cette limite



Caractériser un barcode gap



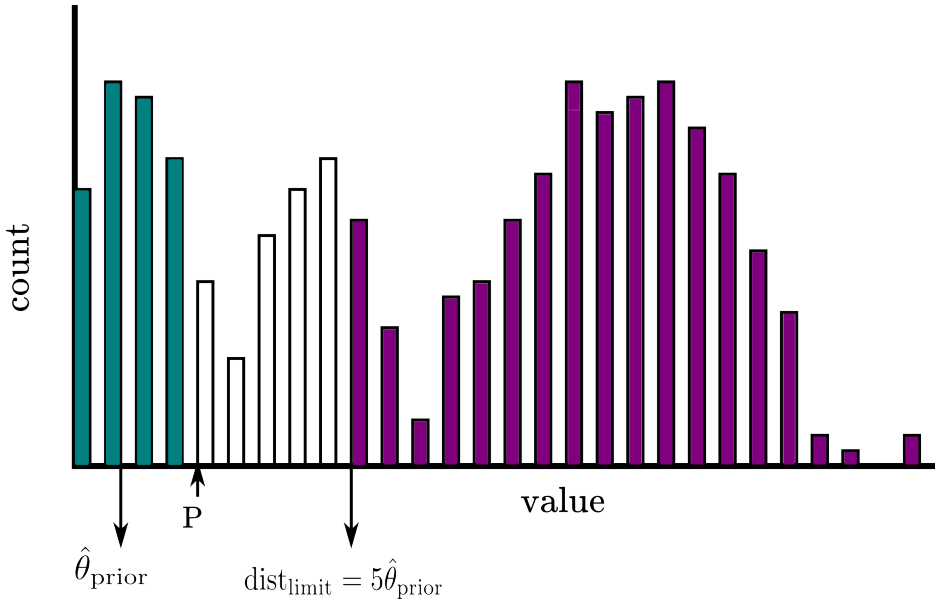
- La localisation du gap est celle du premier maximum qui ne bouge pas pour trois tailles de fenêtres successives.

Distance maximale d'un barcode gap intra-population

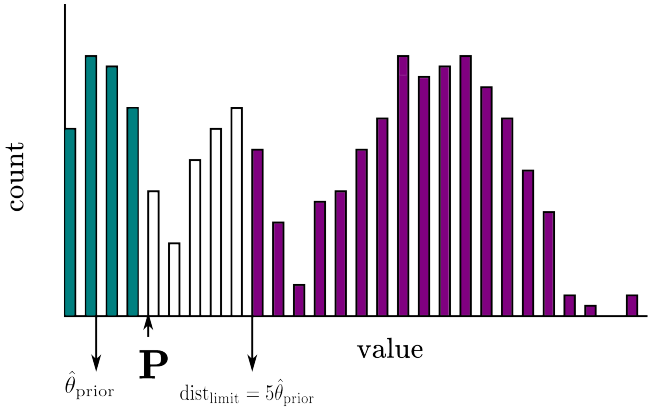
- Pour chaque (n, θ, w), simulation de 10^4 populations panmictiques.
- $\text{dist}_{\text{limit}}$ est déterminé pour que 95% des simulations n'aient aucun gap après $\text{dist}_{\text{limit}}$.

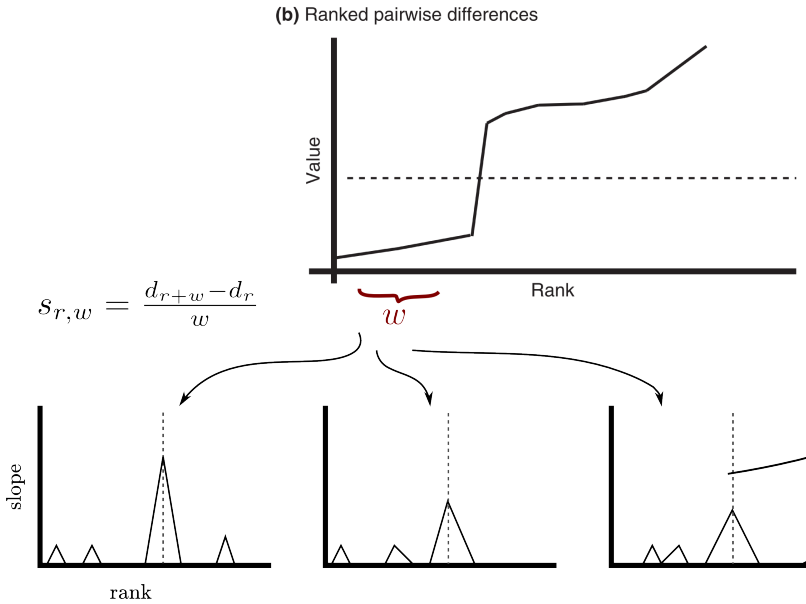
$$\text{dist}_{\text{limit}} = f(n, \theta, w) = a\theta = 2.581\ \theta$$

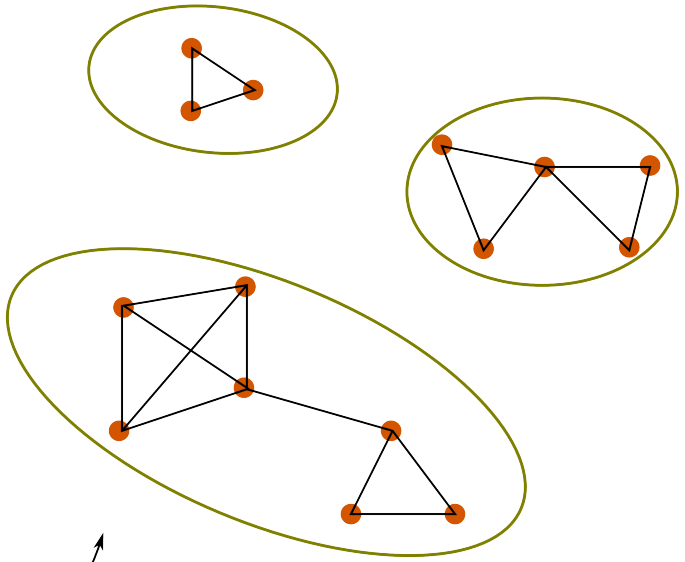
- L'utilisateur doit fixer la limite de diversité intraspécifique : P.
- Estimation de θ par $\hat{\theta}_{\text{prior}}$, la moyenne des différences 2 à 2 inférieures à P.
- Par sécurité, on fixe $\text{dist}_{\text{limit}} = 2a\hat{\theta}_{\text{prior}} \approx 5\hat{\theta}_{\text{prior}}$.



Méthode complète







Potocole d'évaluation

- Un grand nombre de scénarios imaginés.
  - A une seule grande population, hypothèse H0.
  - B 30 espèces, 5 individus chacune, différentes N_e.
  - C diversification avec échantillonnage incomplet (C1), ou naissance-mort (C2).
  - D une population en augmentation (D1) ou diminution (D2).
  - E différentes tailles de populations selon les espèces.
  - F différents nombre d'individus par espèce.
  - G structure géographique au sein des espèces.
  - H effet de la reconstruction de l'arbre de gènes.

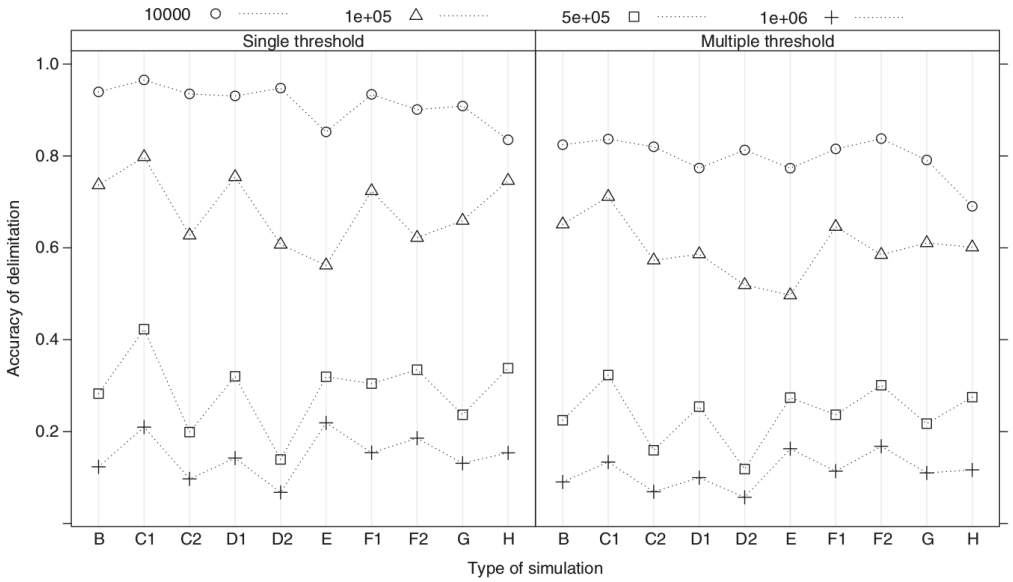
- Deux jeux de données empiriques :

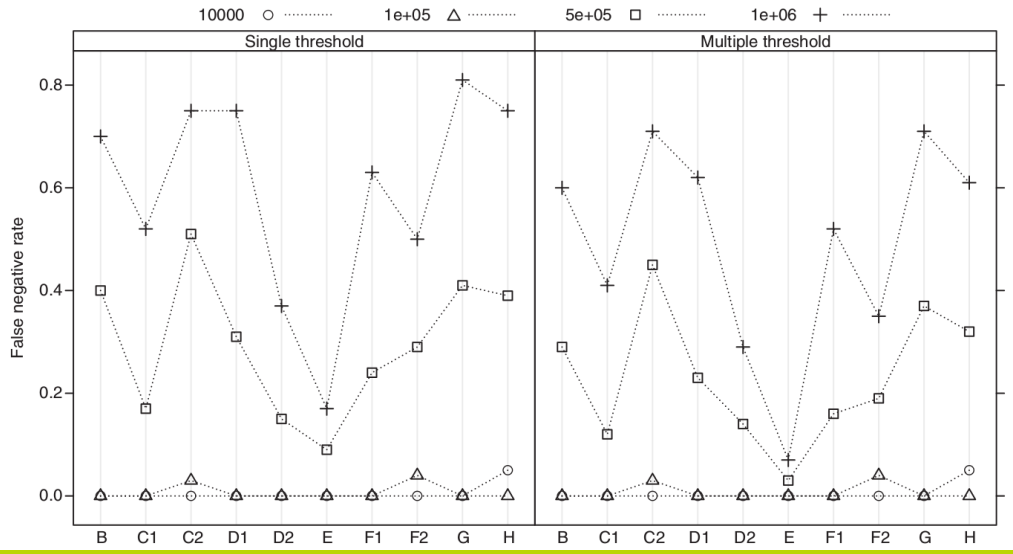
Rivacindela 468 individus, Australie.

Neocicindela 161 individus, Nouvelle Zélande.

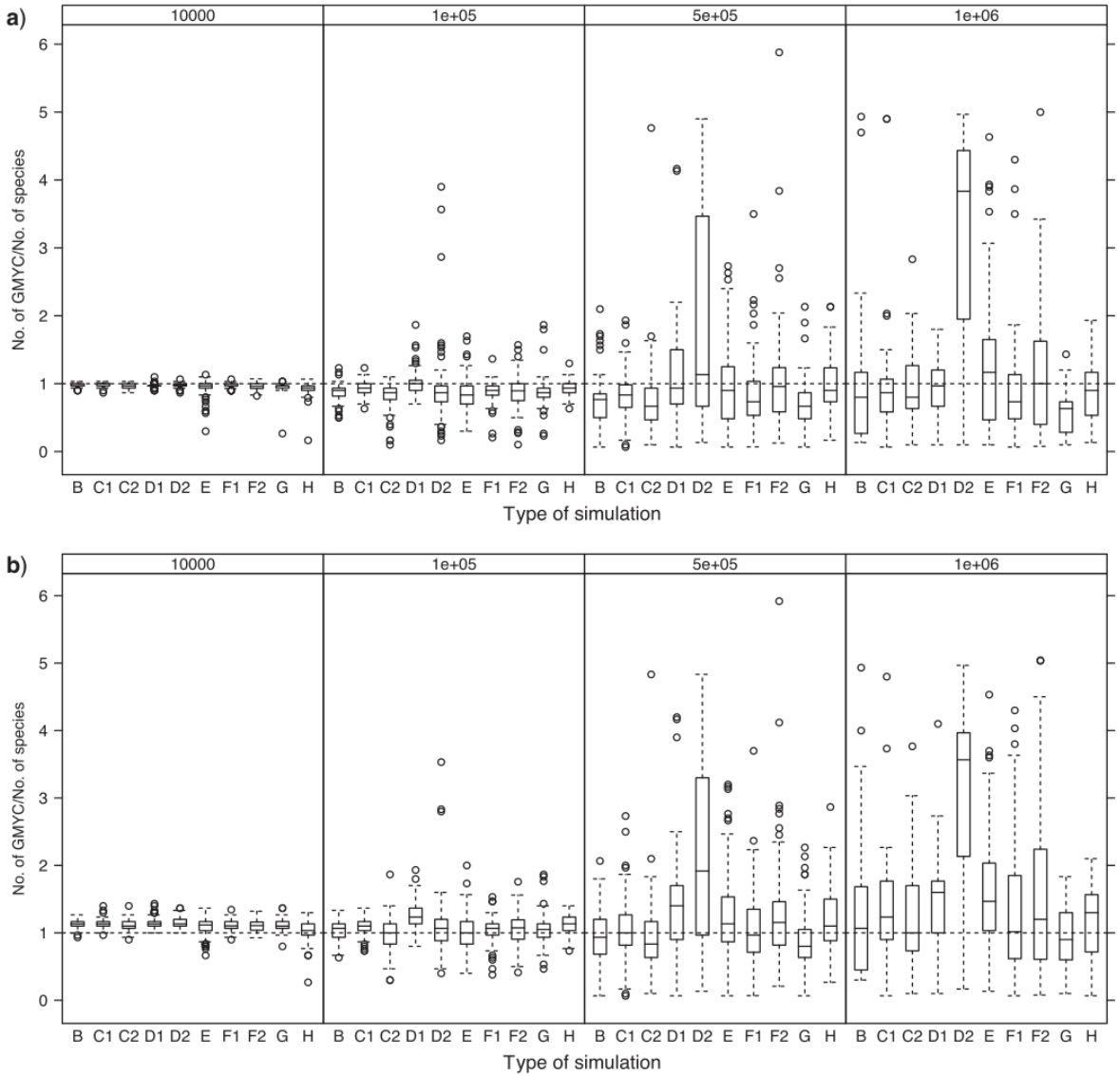
Figure : Tiger Beetle (crédit Muhammad Mahdi Karim)

L'erreur augmente avec N_e

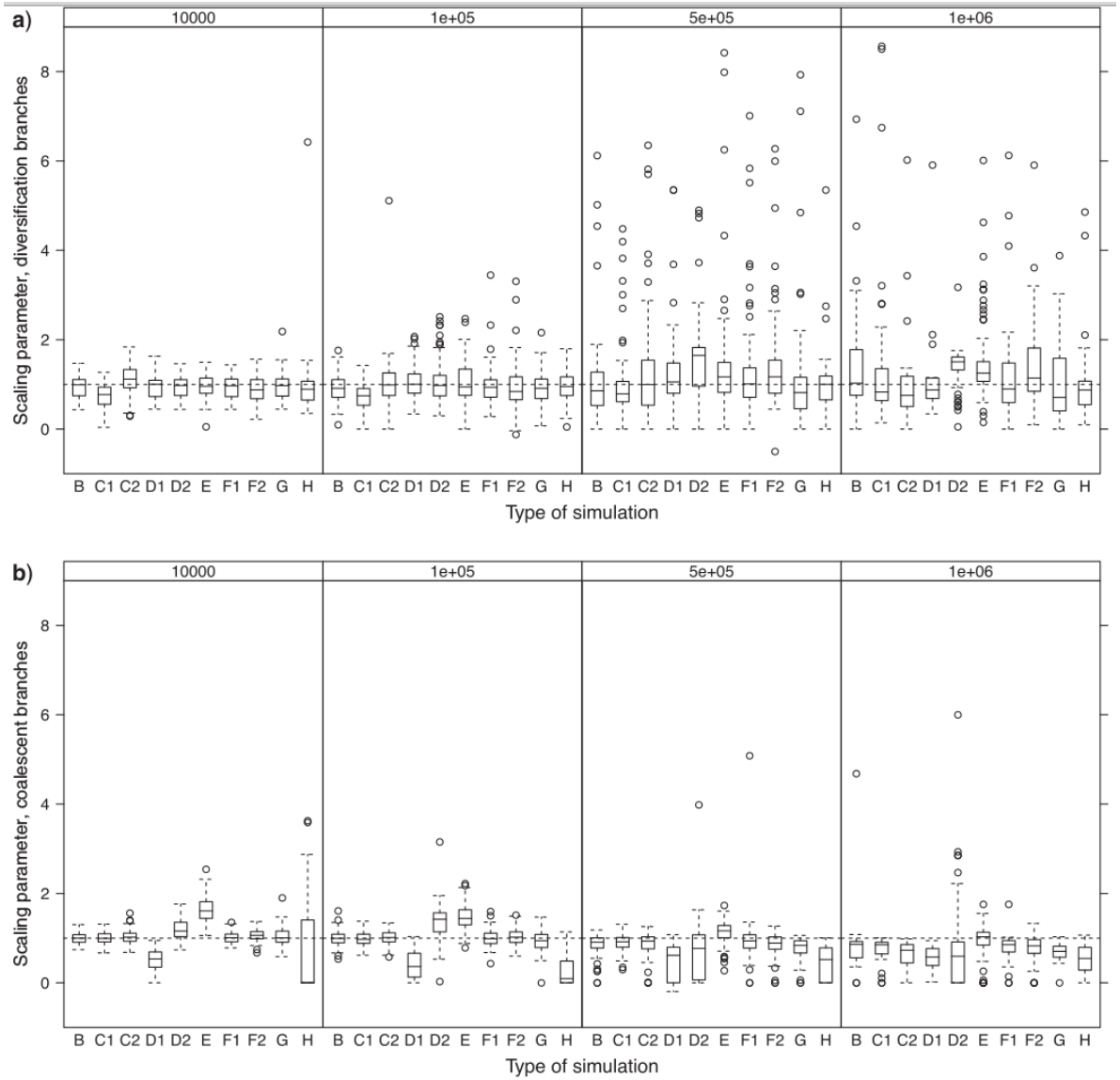




Nombre d'espèces délimitées



Estimation de p_{spec} et p_{coal}

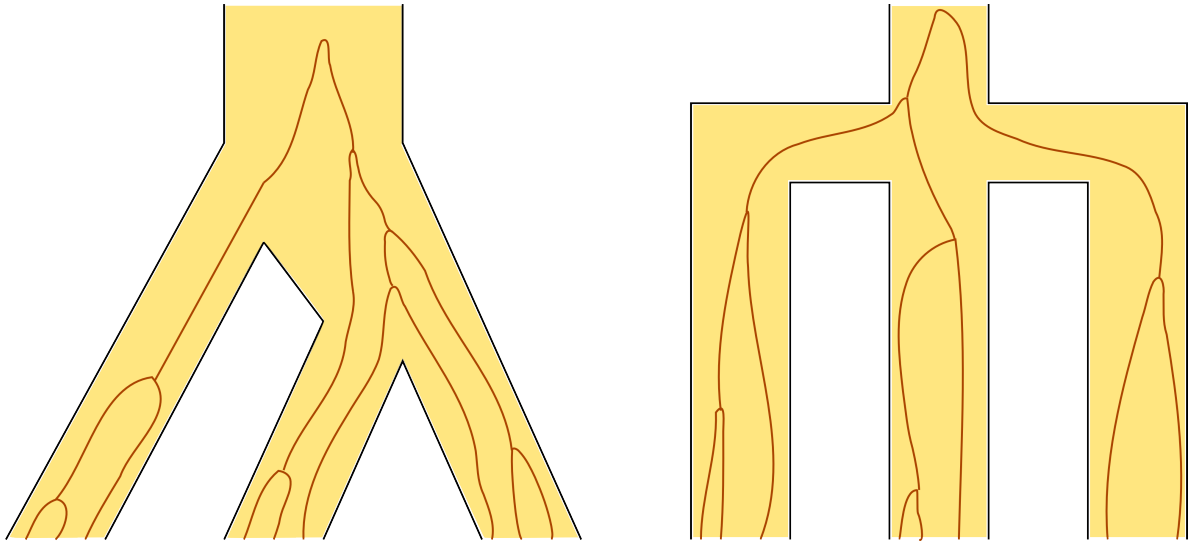


Protocole d'évaluation

- Diversification : 4 scenarios

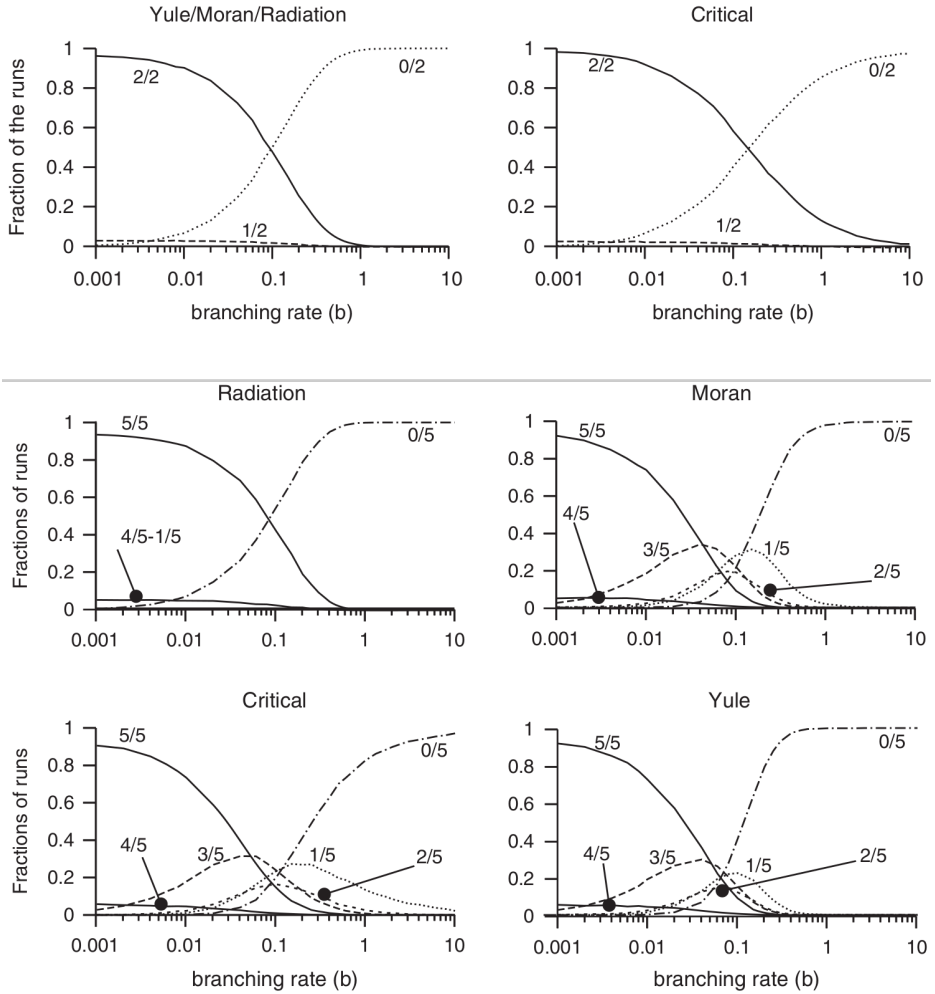
  radiation : un unique événement de spéciation survient à taux b.
  Moran : Kingman à taux $2b/n_s$.
  Yule : naissance à taux b.
  Critique : naissance à taux b, mort à taux b.

- Intra-pop : coalescent de Kingman.
- 6 jeux de données empiriques.



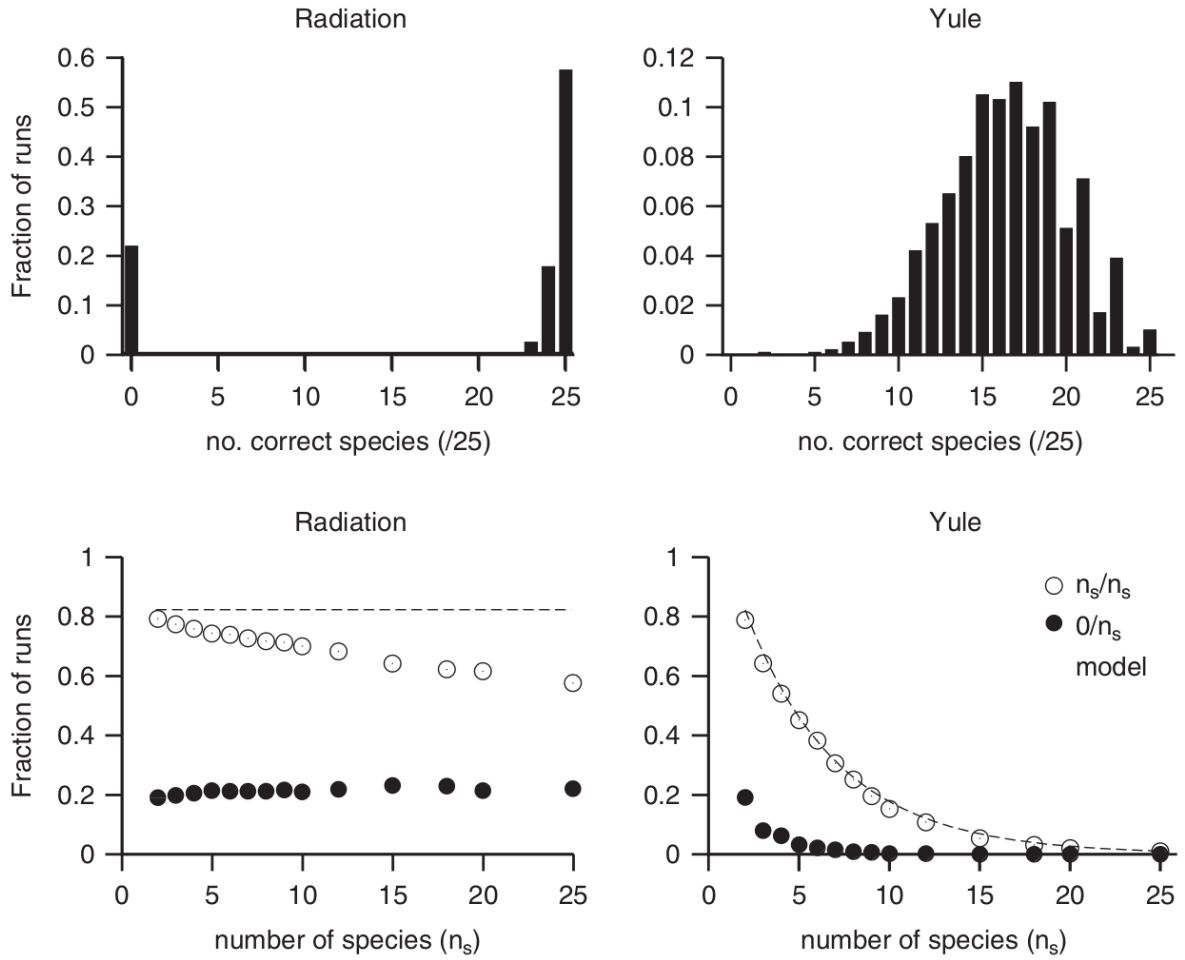
Sur les simulations

- Bonne performance pour b faible.
- Peu d'erreurs d'"oversplitting".



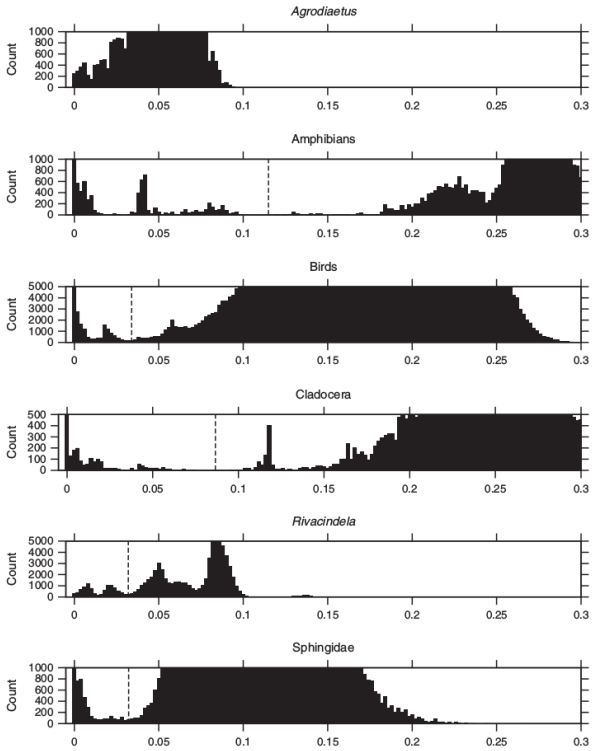
Sur les simulations

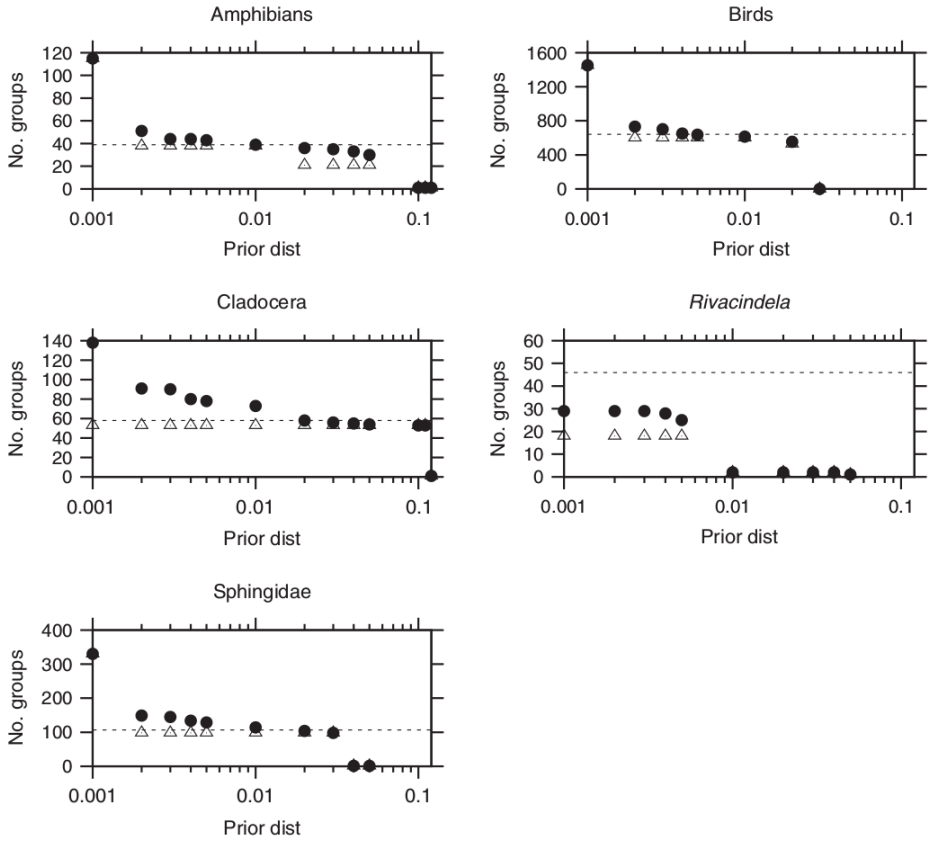
- Assez bonne performance même avec plus d'espèces.



Sur les vraies données

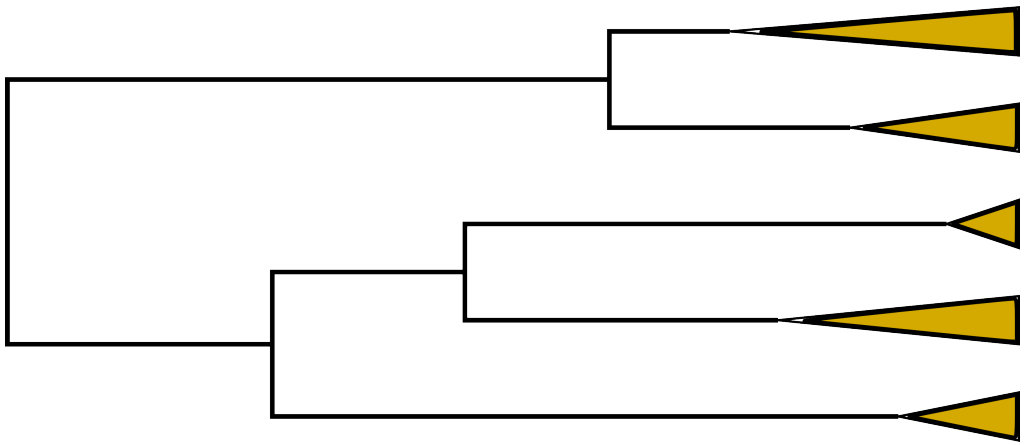
- 1 sur 6 n'a pas de barcode gap : aucun découpage en espèces.
- Les autres sont découpés comme dans les publis précédentes.





Petit bilan

- Une méthode sur arbre reconstruit / une méthode sur séquences.
- Une méthode lente / une méthode rapide.
- Permettent d'obtenir des premières hypothèses d'espèces.
- Tout fonctionne quand la divergence intra-population est toujours plus faible que la divergence inter-espèces, en absence de tri incomplet des lignées.



- Quel gain avec des méthodes prenant en compte plus de loci ?
- Quelles possibilités pour prendre en compte d'autres caractères, non moléculaires ?

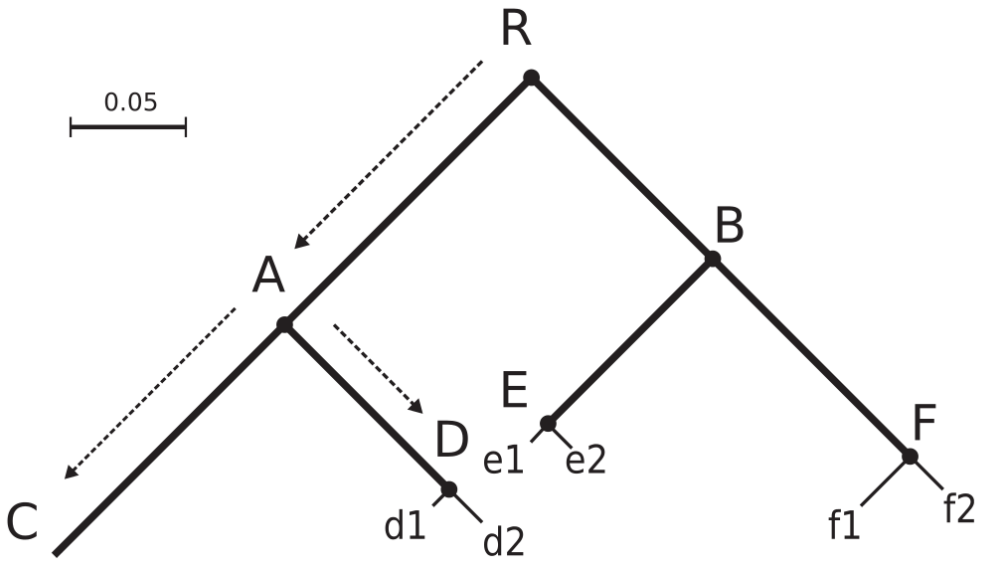
Poisson Tree processes

Le modèle

Principe

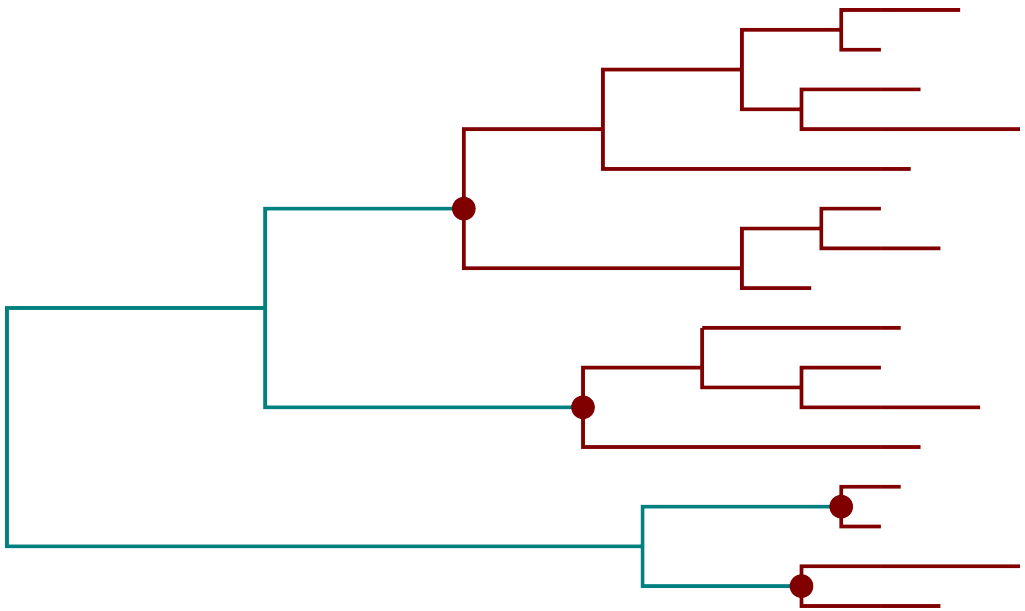
- Prend en entrée un arbre non-ultramétrique.
- Un taux λ_s (naissance d'espèce), shifte en taux λ_c de coalescence intra-pop au niveau de certains noeuds.
- Quand on connait ces noeuds, ils donnent la vraisemblance :

$$\ln \mathcal{L} = \sum_{i=1}^{k} \log(\lambda_s e^{-\lambda_s x_i}) + \sum_{i=k+1}^{n} \log(\lambda_c e^{-\lambda_c x_i})$$



Optimisation de l'assignement des noeuds

- correspond au même problème que pour GMYC.
- 3 "heuristiques" peu compréhensibles.



Utilisation conjointe avec "EPA"

- Utilisent au préablable une méthode de placement des séquences sur un arbre de référence fixe.
- Chaque groupe de séquences qui est greffé à un endroit subit le reste du traitement PTP.