La forme des arbres phylogénétiques

Mémoire de première année

Michel Pain

sous la direction d'Amaury Lambert et de Guillaume Achaz

19 juin 2013

Table des matières

Introduction

Les arbres phylogénétiques sont des arbres représentant les liens de parentés entre les espèces. Dans la classification classique (dite *linnéenne*), les espèces étaient rangées dans les taxons hiérarchisés genre, famille, ordre, classe, etc selon des traits morphologiques ou des caractéristiques en commun (un *taxon* est n'importe quel groupe d'espèces auquel on donne un nom).

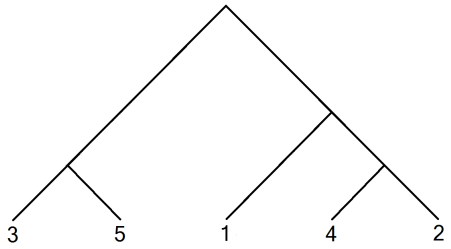
L'étude des "arbres de la vie" était initialement assez subjective, car elle reposait sur le choix des critères morphologiques permettant de dire que deux espèces sont proches. Pour être plus rigoureux, il faut chercher à repérer les évènements de *spéciations*, lors desquels une espèce se sépare en deux. Un *arbre phylogénétique* est alors obtenu en associant à chaque spéciation un nœud et à chaque espèce actuelle une feuille. On obtient ainsi un arbre binaire, dans lequel les espèces sont disposées en clades : un *clade* est un groupe formé de toutes les espèces descendants d'un même ancêtre commun. Ainsi, dans l'arbre, un clade est l'ensemble des feuilles descendant d'un même nœud. Il semble naturel de vouloir classer les espèces en clades, mais dans la classification linnéenne, certains taxons n'étaient pas des clades (par exemple les poissons) d'où l'intérêt d'essayer d'être plus rigoureux dans la définition des arbres phylogénétiques.

Dans ce mémoire on s'intéressera plus particulièrement à la forme des arbres phylogénétiques, c'est-à-dire uniquement à la topologie de l'arbre sans tenir compte de l'échelle temporelle donc de la longueur des branches et de la position relative des nœuds. Dans le cas où on a enlevé l'échelle temporelle d'un arbre phylogénétique, on parle alors plutôt de *cladogramme*.

Ce n'est qu'au milieu du XX$^{\text{ème}}$ siècle que Willi Hennig proposa pour la première fois une méthode clairement définie pour construire un arbre phylogénétique. Le principe est de partir d'un groupe d'espèces et de la donnée de caractères communs à certaines de ces espèces. Dans le tableau suivant, on considère les espèces 1, 2, 3, 4 et 5 et les caractères A, B, C et D. Chaque caractère peut être dans deux états : l'état 0 est l'état ancestral et l'état 1 est l'état dérivé, apparu suite à une mutation.

	Caractères			
Espèces	A	B	C	D
1	0	1	0	0
2	1	1	0	0
3	0	0	1	1
4	1	1	0	0
5	0	0	1	0

Hennig part du principe que chaque caractère nous informe sur un clade, c'est-à-dire que toutes les espèces ayant la version 1 d'un caractère forment un clade : le clade des descendants du premier mutant ayant la version 1 du caractère. Cela présuppose que la mutation de la version 0 à la version 1 n'a eu lieu qu'une seule fois, et qu'il n'y a pas eu de mutation de la version 1 vers la version 0. Sous ces hypothèses, le caractère A montre que $\{2, 4\}$ est un clade et les autres donnent les clades $\{1, 2, 4\}$ et $\{3, 5\}$ (le fait que $\{3\}$ est un clade est évident et donc le caractère D n'apporte pas d'information). On en déduit le cladogramme suivant pour ces 5 espèces :



Cependant, cette méthode a ses limites : si l'on dispose des informations fournies par le tableau suivant qui fait intervenir également les caractères E et F, alors on obtient que $\{1, 4, 5\}$ est une clade, ce qui n'est pas compatible avec les autres données.

	Caractères					
Espèces	A	B	C	D	E	F
1	0	1	0	0	1	1
2	1	1	0	0	0	0
3	0	0	1	1	0	0
4	1	1	0	0	1	1
5	0	0	1	0	1	1

Pour résoudre une telle incompatibilité, il existe deux méthodes. La méthode par parcimonie autorise qu'il y ait pour un caractère des mutations de la version 1 vers la version 0 et plus d'une mutation de 0 vers 1 et cherche alors l'arbre phylogénétique qui minimise le nombre de tels évènements. La méthode par compatibilité n'autorise pas de tels évènements, cherche le plus grand ensemble de caractères compatibles et construit l'arbre à partir de ces caractères seulement.

Les débuts du séquençage des protéines permit d'obtenir des données plus précises que celles obtenues en comparant les caractères morphologiques : on cherche les différences de nucléotides entre les différentes versions de la protéine dans plusieurs espèces. Cet outil rendit les méthodes précédentes plus objectives.

D'autres méthodes ont également été développées pour construire un arbre à partir de telles données. Les méthodes de matrices de distance consistent à associer aux données une matrice A contenant la distance entre les espèces (par exemple, le nombre de nucléotides distincts entre les séquences de la protéine pour deux espèces) et à un arbre une matrice B contenant la distance entre les espèces dans l'arbre (la longueur du plus court chemin entre deux feuilles) et à chercher l'arbre pour lequel la matrice A et la plus proche de la matrice B (il existe plusieurs définitions pour la distance entre A et B). Les méthodes de maximum de vraisemblance consistent à partir d'un modèle régissant les mutations du code génétique et de données regroupant des séquences de protéines pour les différentes espèces et à évaluer pour chaque arbre, la probabilité que les mutations dictées par l'arbre aient lieu sous le modèle. On conserve alors l'arbre qui maximise cette probabilité.

Aujourd'hui, avec la possibilité de séquencer le génome, on obtient toujours plus de données pour construire les arbres phylogénétiques, en comparant des séquences de génomes très conservées chez différentes espèces. Il existe une base de donnée, TreeBASE (`www.treebase.org`), qui regroupe des milliers d'arbres phylogénétiques ([4]), certains ordonnant plus de 400 espèces ([2]).

L'objectif de ce mémoire est de chercher un modèle probabiliste simple générant des cladogrammes aléatoires ayant les mêmes propriétés topologiques que les arbres obtenus par les études biologiques. Nous n'étudierons donc pas plus les méthodes de construction des arbres et nous contenterons d'utiliser les arbres disponibles dans TreeBASE pour comparer les modèles aux données.

Nous commencerons par étudier une famille de distributions à un paramètre sur les partitions, pour définir par analogie les propriétés que l'on souhaite vérifiées par une distribution sur les cladogrammes. Nous construirons ensuite des premiers modèles à partir de ces propriétés, puis introduirons une famille de modèles à un paramètre qui contient les modèles précédents. Nous verrons alors comment l'une des distributions de cette famille s'approche particulièrement bien des données.

1 Axiomatisation des propriétés des cladogrammes aléatoires

On cherche à définir les propriétés "naturelles" que doit vérifier une distribution sur l'ensemble des cladogrammes à n feuilles.

1.1 Distribution sur les partitions

Pour trouver les propriétés que l'on souhaite vérifiées par une distribution sur les cladogrammes, on s'intéresse tout d'abord aux distributions sur les partitions.

On note $\mathcal{P}_n$ l'ensemble des partitions (non-ordonnées) de $[\![1;n]\!]$.

Définition 1

On définit la famille $(P_\theta^{(n)})_{n\geq 1}$ paramétrée par $\theta > 0$ de distributions sur les $\mathcal{P}_n$ par :

$$P_\theta^{(n)}(\{A_1,\dots,A_k\}) = \frac{\Gamma(\theta)}{\Gamma(n+\theta)}\prod_{i=1}^{n}((i-1)!\ \theta)^{m_i}$$

où $m_i = \#\{j \in [\![1;k]\!] \mid \#(A_j) = i\}$ pour $\{A_1,\dots,A_k\} \in \mathcal{P}_n$.

Démonstration : Il faut vérifier que, pour tous $\theta > 0$ et $n \geq 1$, $P_\theta^{(n)}$ est bien de masse totale 1. Soit $\theta > 0$, on montre le résultat par récurrence sur n :

- $n = 0$: $\mathcal{P}_1 = \Big\{\big\{\{1\}\big\}\Big\}$ et $P_\theta^{(1)}(\big\{\{1\}\big\}) = \frac{1}{\theta}\cdot\theta = 1$.
- $n-1 \to n$: Se donner une partition de $[\![1;n]\!]$ revient à se donner une partition $\{A_1,\dots,A_k\}$ de $[\![1;n-1]\!]$ et à ensuite soit rajouter n à l'un des A_j, soit rajouter le singleton $\{n\}$ comme sous-ensemble de la partition, on a donc, avec des unions disjointes :

$$\mathcal{P}_n = \bigsqcup_{\{A_1,\dots,A_k\}\in\mathcal{P}_{n-1}} \left(\Big\{\big\{A_1,\dots,A_k,\{n\}\big\}\Big\} \sqcup \bigsqcup_{j=1}^{k}\Big\{\big\{A_1,\dots,A_j\cup\{n\},\dots,A_k\big\}\Big\}\right).$$

On remarque d'autre part que :

$$P_\theta^{(n)}(\{A_1,\dots,A_k\}) = \frac{\Gamma(\theta)}{\Gamma(n+\theta)}\prod_{j=1}^{k}((\#A_j-1)!\;\theta).$$

La masse totale de $P_\theta^{(n)}$ est donc :

$$\begin{aligned}
P_\theta^{(n)}(\mathcal{P}_n) &= \sum_{\{A'_1,\dots,A'_k\}\in\mathcal{P}_n} P_\theta^{(n)}(\{A'_1,\dots,A'_k\}) \\
&= \sum_{\{A_1,\dots,A_k\}\in\mathcal{P}_{n-1}} \Bigg[P_\theta^{(n)}\Big(\Big\{\{A_1,\dots,A_k,\{n\}\}\Big\}\Big) \\
&\qquad + \sum_{j=1}^{k} P_\theta^{(n)}\Big(\Big\{\{A_1,\dots,A_j\cup\{n\},\dots,A_k\}\Big\}\Big)\Bigg] \\
&= \sum_{\{A_1,\dots,A_k\}\in\mathcal{P}_{n-1}} \Bigg[\frac{1}{n+\theta-1}\cdot\theta\cdot P_\theta^{(n-1)}\Big(\Big\{\{A_1,\dots,A_k\}\Big\}\Big) \\
&\qquad + \sum_{j=1}^{k} \frac{1}{n+\theta-1}\cdot\frac{(\#A_j)!\;\theta}{(\#A_j-1)!\;\theta}\cdot P_\theta^{(n-1)}\Big(\Big\{\{A_1,\dots,A_k\}\Big\}\Big)\Bigg] \\
&= \sum_{\{A_1,\dots,A_k\}\in\mathcal{P}_{n-1}} \Bigg[P_\theta^{(n-1)}\Big(\Big\{\{A_1,\dots,A_k\}\Big\}\Big)\cdot\frac{1}{n+\theta-1}\cdot\Big(\theta+\sum_{j=1}^{k}\#A_j\Big)\Bigg] \\
&= \sum_{\{A_1,\dots,A_k\}\in\mathcal{P}_{n-1}} P_\theta^{(n-1)}\Big(\Big\{\{A_1,\dots,A_k\}\Big\}\Big) \\
&= 1.
\end{aligned}$$

■

Pour chaque $\theta>0$, $(P_\theta^{(n)})_{n\geq 1}$ vérifie les propriétés "naturelles" que nous allons définir (cette famille est même caractérisée par ces propriétés comme l'énonce la proposition 1).

Définition 2

On définit pour une famille $(P^{(n)})_{n\geq 1}$ de distributions sur les $\mathcal{P}_n$ les propriétés :

(i) *Échangeabilité* : Pour tout $n\geq 1$, la distribution est invariante par permutation des étiquettes $\{1,\dots,n\}$:
$\forall\{A_1,\dots,A_k\}\in\mathcal{P}_n, \forall\sigma\in\mathcal{S}_n, P^{(n)}(\{A_1,\dots,A_k\}) = P^{(n)}(\{A_1^\sigma,\dots,A_k^\sigma\})$ où $A_j^\sigma=\{\sigma(a)\mid a\in A_j\}$.

(ii) *Invariance par échantillonnage* : Pour tout $n\geq 1$, $P^{(n)}$ induit une distribution sur $\mathcal{P}_{n-1}$ par l'action de supprimer n et cette distribution est $P^{(n-1)}$:
En notant $\chi_n:\mathcal{P}_n\longrightarrow\mathcal{P}_{n-1}$, qui à une partition de $[\![1;n]\!]$ associe la partition de $[\![1;n-1]\!]$ obtenue en supprimant n (et en supprimant $\varnothing$ s'il apparaît alors dans la liste des sous-ensembles), alors on a pour $\{A_1,\dots,A_k\}\in\mathcal{P}_{n-1}$ on a $P^{(n)}\big(\chi_n^{-1}(\{A_1,\dots,A_k\})\big) = P^{(n-1)}(\{A_1,\dots,A_k\})$.

(iii) *Suppression d'un sous-ensemble* : Pour tous $1\leq j<n$, sachant que la partition $P^{(n)}$ contient le sous-ensemble $[\![j+1;n]\!]$, la partition obtenue en supprimant

$[\![j+1;n]\!]$ est distribuée selon $P^{(j)}$:
En notant $\mathcal{A}_{n,j} = \Big\{\{A_1,\dots,A_k\} \in \mathcal{P}_n \mid \exists j \in [\![1;k]\!] : A_j = [\![j+1;n]\!]\Big\}$, et $\phi_{n,j} : \{A_1,\dots,A_k\} \in \mathcal{A}_{n,j} \longmapsto \{A_1,\dots,A_k\} \setminus \{[\![j+1;n]\!]\} \in \mathcal{P}_j$, alors pour $\{A_1,\dots,A_k\} \in \mathcal{P}_j$, $P^{(n)}(\phi_{n,j}^{-1}(\{A_1,\dots,A_k\}))/P^{(n)}(\mathcal{A}_{n,j}) = P^{(j)}(\{A_1,\dots,A_k\})$.

Proposition 3

Soit $(P^{(n)})_{n\geq 1}$ une famille de distributions sur les $\mathcal{P}_n$ telle que $P^{(2)}$ ne soit pas une mesure de Dirac, on a l'équivalence :

$$(P^{(n)})_{n\geq 1}\text{vérifie (i), (ii) et (iii)} \Leftrightarrow \exists \theta > 0 : (P^{(n)})_{n\geq 1} = (P_\theta^{(n)})_{n\geq 1}.$$

Démonstration :

- Soit $\theta > 0$, montrons que $(P_\theta^{(n)})_{n\geq 1}$ vérifie (i), (ii) et (iii).
  (i) Trivial car $P_\theta^{(n)}(\{A_1,\dots,A_k\})$ ne dépend que des cardinaux des A_j.
  (ii) Soit $\{A_1,\dots,A_k\} \in \mathcal{P}_{n-1}$, on remarque que :

$$\chi_n^{-1}(\{A_1,\dots,A_k\}) = \Big\{\{A_1,\dots,A_k,\{n\}\}\Big\} \sqcup \bigsqcup_{j=1}^{k} \Big\{\{A_1,\dots,A_j\cup\{n\},\dots,A_k\}\Big\}.$$

  On a donc :

$$\frac{P^{(n)}\big(\chi_n^{-1}(\{A_1,\dots,A_k\})\big)}{P^{(n-1)}(\{A_1,\dots,A_k\})} = \frac{1}{n+\theta-1}\cdot\Big(\theta + \sum_{j=1}^{k}\frac{(\#A_j)!\,\theta}{(\#A_j-1)!\,\theta}\Big) = 1.$$

  (iii) $\phi_{n,j} : \mathcal{A}_{n,j} \longrightarrow \mathcal{P}_j$ est une bijection et on a donc :

$$\mathcal{A}_{n,j} = \bigsqcup_{\{A_1,\dots,A_k\}\in\mathcal{P}_j} \Big\{\{A_1,\dots,A_k,[\![j+1;n]\!]\}\Big\}.$$

  On en déduit :

$$\begin{aligned}
P_\theta^{(n)}(\mathcal{A}_{n,j}) &= \sum_{\{A_1,\dots,A_k\}\in\mathcal{P}_j} P_\theta^{(n)}\Big(\Big\{\{A_1,\dots,A_k,[\![j+1;n]\!]\}\Big\}\Big)\\
&= \sum_{\{A_1,\dots,A_k\}\in\mathcal{P}_j}\Big[P_\theta^{(j)}\Big(\Big\{\{A_1,\dots,A_k\}\Big\}\Big)\cdot\frac{\Gamma(j+\theta)}{\Gamma(n+\theta)}\cdot(n-j-1)!\,\theta\Big]\\
&= \frac{\Gamma(j+\theta)}{\Gamma(n+\theta)}\cdot(n-j-1)!\,\theta\\
&= \frac{P_\theta^{(n)}\Big(\Big\{\{A_1,\dots,A_k,[\![j+1;n]\!]\}\Big\}\Big)}{P_\theta^{(j)}\Big(\Big\{\{A_1,\dots,A_k\}\Big\}\Big)}.
\end{aligned}$$

- Soit $(P^{(n)})_{n\geq 1}$ une famille de distributions sur les $\mathcal{P}_n$ qui vérifie (i), (ii) et (iii).
  * $\mathcal{P}_1$ est un singleton donc $P^{(1)} = P_\theta^{(1)}$ pour tout $\theta > 0$.
  * $\mathcal{P}_2 = \Big\{\{\{1\},\{2\}\},\{\{1,2\}\}\Big\}$ et pour $\theta > 0$,

$$P_\theta^{(2)}(\{\{1\},\{2\}\}) = \frac{\theta}{\theta+1} \text{ et } P_\theta^{(2)}(\{\{1,2\}\}) = \frac{1}{\theta+1}.$$

Or $\theta \longmapsto \frac{\theta}{\theta+1}$ est une bijection de $\mathbb{R}_+^* \longrightarrow]0;1[$ et $P^{(2)}\big(\{\{1\},\{2\}\}\big) \in]0;1[$ car $P^{(2)}$ n'est pas un Dirac, donc il existe $\theta > 0$ tel que $P^{(2)} = P_\theta^{(2)}$.

∗ Raisonnons par récurrence sur $n \geq 2$ avec le θ fixé pour $n = 2$. Soit $n \geq 3$, supposons que $\forall j \in [\![1, n-1]\!], P^{(j)} = P_\theta^{(j)}$ et montrons que $P^{(n)} = P_\theta^{(n)}$.
Commençons par montrer que $P^{(n)}$ et $P_\theta^{(n)}$ coïncident sur $\mathcal{A}_{n,n-1}$. D'après (iii),

$$\forall \{A_1, \dots, A_k\} \in \mathcal{P}_{n-1}, \frac{P^{(n)}\big(\{A_1, \dots, A_k, \{n\}\}\big)}{P^{(n)}(\mathcal{A}_{n,n-1})} = P^{(j)}\big(\{A_1, \dots, A_k\}\big).$$

Donc, comme $P^{(n-1)} = P_\theta^{(n-1)}$, on connaît $P^{(n)}$ sur $\mathcal{A}_{n,n-1}$ à une constante multiplicative près : pour $\{A_1, \dots, A_k\} \in \mathcal{P}_{n-1}$,

$$\begin{aligned}\frac{P^{(n)}\big(\{A_1, \dots, A_k, \{n\}\}\big)}{P^{(n)}\big(\{\{1\}, \dots, \{n-1\}, \{n\}\}\big)} &= \frac{P^{(n-1)}\big(\{A_1, \dots, A_k\}\big)}{P^{(n-1)}\big(\{\{1\}, \dots, \{n-1\}\}\big)} \\ &= \frac{P_\theta^{(n-1)}\big(\{A_1, \dots, A_k\}\big)}{P_\theta^{(n-1)}\big(\{\{1\}, \dots, \{n-1\}\}\big)} \\ &= \frac{P_\theta^{(n)}\big(\{A_1, \dots, A_k, \{n\}\}\big)}{P_\theta^{(n)}\big(\{\{1\}, \dots, \{n-1\}, \{n\}\}\big)}.\end{aligned}$$

Il suffit donc de déterminer $a := P^{(n)}\big(\{\{1\}, \dots, \{n\}\}\big)$. On utilise (ii) :

$$P^{(n-1)}\big(\{\{1\}, \dots, \{n-1\}\}\big) = a + \sum_{j=1}^{n-1} P^{(n)}\big(\{\{1\}, \dots, \{j, n\}, \dots, \{n-1\}\}\big).$$

Or, d'après (i), pour tout $j \in [\![1, n-1]\!]$,

$$\begin{aligned}P^{(n)}\big(\{\{1\}, \dots, \{j, n\}, \dots, \{n-1\}\}\big) &= P^{(n)}\big(\{\{1, 2\}, \{3\}, \dots, \{n\}\}\big) \\ &= \frac{P^{(n-1)}\big(\{\{1, 2\}, \{3\}, \dots, \{n-1\}\}\big)}{P^{(n-1)}\big(\{\{1\}, \dots, \{n-1\}\}\big)} \cdot a.\end{aligned}$$

Donc on obtient a uniquement en fonction de $P^{(n-1)}$, et, comme $P_\theta^{(n-1)}$ vérifie aussi (i) et (ii) on obtient une relation identique pour $P_\theta^{(n)}\big(\{\{1\}, \dots, \{n\}\}\big)$. Avec $P^{(n-1)} = P_\theta^{(n-1)}$, on en déduit que $a = P_\theta^{(n)}\big(\{\{1\}, \dots, \{n\}\}\big)$ et donc que $P^{(n)}$ et $P_\theta^{(n)}$ coïncident sur $\mathcal{A}_{n,n-1}$ et donc, d'après (i), sur toutes les partitions contenant au moins un singleton.
Montrons qu'elles coïncident sur $\mathcal{P}_n$ tout entier. Soit $\{A_1, \dots, A_k\} \in \mathcal{P}_n$ différent de $\{\{1\}, \dots, \{n\}\}$. Alors $k \geq 2$ et, d'après (i), quitte à permuter les entiers ce qui ne change pas la masse de $\{A_1, \dots, A_k\}$, on peut supposer que A_k est de la forme $[\![j+1, n]\!]$ avec $j \geq 1$ (car $k \geq 2$).
D'après (ii), on a :

$$\frac{P^{(n)}(\{A_1, \dots, A_{k-1}, [\![j+1, n]\!]\})}{P^{(n)}(\mathcal{A}_{n,j})} = P^{(j)}(\{A_1, \dots, A_{k-1}\})$$

$$\text{et } \frac{P^{(n)}(\{\{1\}, \dots, \{j\}, [\![j+1, n]\!]\})}{P^{(n)}(\mathcal{A}_{n,j})} = P^{(j)}(\{\{1\}, \dots, \{j\}\}).$$

On en déduit :

$$P^{(n)}(\{A_1,\dots,A_{k-1},[\![j{+}1,n]\!]\}) = \frac{P^{(j)}(\{A_1,\dots,A_{k-1}\})}{P^{(j)}(\{\{1\},\dots,\{j\}\})}\cdot P^{(n)}(\{\{1\},\dots,\{j\},[\![j{+}1,n]\!]\}).$$

On a la même relation avec $P_\theta^{(n)}$ et $P_\theta^{(j)}$. Or, comme la partition $j \geq 1$, $\{\{1\},\dots,\{j\},[\![j+1,n]\!]\}$ contient au moins un singleton, d'après ce qui précède, $P^{(n)}(\{\{1\},\dots,\{j\},[\![j+1,n]\!]\}) = P_\theta^{(n)}(\{\{1\},\dots,\{j\},[\![j+1,n]\!]\})$, donc on en déduit que $P^{(n)}(\{A_1,\dots,A_k\}) = P_\theta^{(n)}(\{A_1,\dots,A_k\})$.
Donc $P^{(n)}$ et $P_\theta^{(n)}$ coïncident sur $\mathcal{P}_n \setminus \{\{1\},\dots,\{n\}\}$, donc sur $\mathcal{P}_n$ tout entier car elles sont de masse totale 1. ■

Au delà de ces trois propriétés mathématiques, cette famille de distributions vérifie une autre propriété, importante du point de vue biologique : elle peut s'interpréter par un processus temporel simple (ce que nous appellerons la propriété (b) pour les cladogrammes). On n'entrera pas dans le détail ici et on se contentera de décrire grossièrement ce processus.

On se place dans une population de taille N, qui évolue par générations discrètes. À chaque nouvelle génération, chaque individu tire son unique parent uniformément parmi les N individus de la génération précédente. Si à un instant donné on choisit n individus de la population, en prenant ces individus comme feuilles, on obtient un arbre en représentant les lignées de descendance qui mènent à ces individus, jusqu'à ce qu'elles se rejoignent toutes. Quand $N \longrightarrow +\infty$, la hauteur de l'arbre est de l'ordre de N pour n fixé.

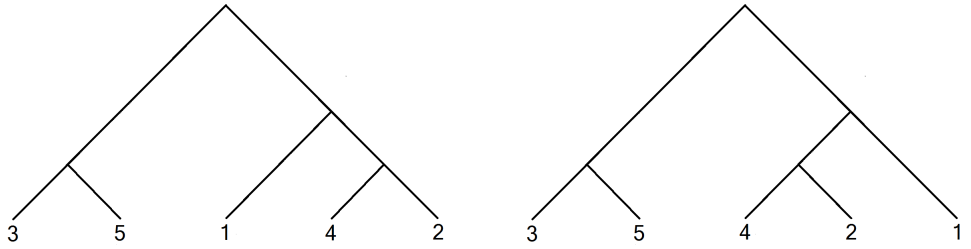
On suppose en outre que des mutations apparaissent sur un gène fixé, chaque mutation faisant apparaître un nouvel allèle. Si on considère l'arbre associé à n individus choisis à un instant donné, les mutations ont lieu le long d'une branche de l'arbre selon un processus de Poisson de paramètre λ/N, de sorte que lorsque $N \longrightarrow +\infty$, comme la hauteur de l'arbre est de l'ordre de N, le nombre de mutations reste fini.

Si l'on choisit uniformément dans la population n individus, qu'on leur donne des étiquettes de 1 à n de manière uniforme et des allèles selon le procédé décrit ci-dessus, alors on peut leur associer une partition de $[\![1,n]\!]$, en les regroupant par allèle commun. La partition aléatoire ainsi obtenue converge en loi quand $N \longrightarrow +\infty$ vers l'un des P_n^θ, où le paramètre θ dépend de λ.

Ce modèle neutre a rencontré un certain succès et a même eu un impact sur la génétique non mathématique. On se demande donc ici s'il existe une théorie neutre analogue pour les arbres phylogénétiques.

1.2 Axiomatisation des distributions sur les cladogrammes

Pour $n \geq 1$, on note $\mathcal{C}_n$ l'ensemble des cladogrammes à n feuilles. Un cladogramme $C \in \mathcal{C}_n$ est un arbre binaire dont les n feuilles représentent des espèces que l'on a étiquetées de 1 à n : si l'on permute deux étiquettes d'espèces, on obtient un autre cladogramme, même s'il a la même forme. En revanche, on ne tient pas compte des orientations gauche et droite de part et d'autre d'un nœud : lors d'une spéciation, les deux nouvelles espèces sont indifférenciées. Ainsi les deux arbres suivants représentent le même cladogramme :



Définition 4

On définit pour une famille $(T_n)_{n\geq 1}$ de distributions sur les $\mathcal{C}_n$ les deux propriétés qui semblent être naturelles :

(i) *Échangeabilité* : Pour tout $n \geq 1$, la distribution est invariante par permutation des étiquettes des n espèces :
Pour $C \in \mathcal{C}_n$ et $\sigma \in \mathcal{S}_n$, si on note C_σ l'arbre obtenu par permutation σ des espèces, alors on a $T_n(C) = T_n(C_\sigma)$.

(ii) *Suppression d'un clade* : Pour tout $1 \leq k < n$, sachant que le cladogramme T_n contient le clade $\{k+1, \ldots, n\}$, le cladogramme obtenu en supprimant ce clade est distribué selon T_k :
On note $\mathcal{A}_{n,k}$ l'ensemble des $C \in \mathcal{C}_n$ tels qu'il existe un nœud dans C dont l'ensemble des descendants est exactement $\{k+1, \ldots, n\}$. Pour $C \in \mathcal{C}_n$, on note $\varphi_{n,k}(C) \in \mathcal{C}_k$ le cladogramme obtenu en supprimant ce nœud et tous ses descendants. Alors on a pour $C \in \mathcal{C}_k$, $T_n(\varphi_{n,k}^{-1}(C))/T_n(\mathcal{A}_{n,k}) = T_k(C)$.

En particulier, la propriété (i) signifie que la probabilité d'un cladogramme dépend uniquement de la forme de l'arbre et pas de la disposition des étiquettes $\{1, \ldots, n\}$ sur ses feuilles. On pourra donc représenter un cladogramme sans étiquettes, puisque tous les cladogrammes ayant cette forme et une disposition quelconque des étiquettes ont même probabilité.

Par analogie avec les distributions sur les partitions, on cherche une famille à un paramètre de distributions sur les $\mathcal{C}_n$ telle que :

(a) Une suite $(T_n)_{n\geq 1}$ de distributions appartient à la famille si et seulement si elle vérifie (i) et (ii).

(b) Ces cladogrammes aléatoires peuvent être décrit à partir d'un modèle d'évolution des espèces au cours du temps.

La propriété (b) traduit le fait que la distribution n'est pas uniquement une famille mathématique qui convient, mais qu'elle a aussi une signification biologique.

2 Le modèle ERM

On cherche à définir une telle famille à un paramètre de distributions sur les $\mathcal{C}_n$ en partant de (b).

2.1 Le procédé de Yule

En 1924, avant l'utilisation des arbres phylogénétiques, les espèces étaient hiérarchisées selon la classification linnéenne. Le *genre* est le taxon directement au-dessus de l'espèce dans cette classification, il est composé d'espèces relativement proches, ce qui laisse une certaine part à la subjectivité. À partir de ces données, on peut s'intéresser au nombre d'espèces par genres et remarquer que le nombre d'espèces le plus fréquent est 1 et que la distribution est à queue lourde.

Yule chercha alors une famille de distributions à un paramètre à laquelle on pourrait comparer les données, où le paramètre représenterait la part de subjectivité dans la définition de "relativement proches". Dans [7], il obtient ces distributions à partir de (b), c'est-à-dire à partir d'un modèle crédible de l'évolution.

Définition 5

Une *population* est une suite de variables aléatoires $(X_t)_{t\geq 0}$ à valeurs dans $\mathbb{N}$. On dit qu'une population $(X_t)_{t\geq 0}$ suit le *procédé de Yule* ou *procédé de naissance linéaire* de paramètre λ si elle vérifie :

- La population commence au temps 0 avec un individu.
- Au fil du temps, les individus peuvent donner naissance à un nouvel individu, la probabilité qu'un individu particulier donne naissance pendant $[t, t+dt]$ étant λdt.

En d'autres termes, cela signifie qu'il existe $(\tau_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes, telle que, pour tout $n \geq 1$, τ_n suit la loi exponentielle de paramètre $n\lambda$ et $X_t = \max\{n \geq 0 \mid \tau_1 + \cdots + \tau_n \leq t\}$.

Dans cette définition, τ_n représente le temps entre la $(n-1)^{\text{ème}}$ et la $n^{\text{ème}}$ naissance et donc $\tau_1 + \cdots + \tau_n$ représente l'instant de la $n^{\text{ème}}$ naissance. À cet instant, il y a alors $n+1$ individus dans la population.

Proposition 6

Si $(X_t)_{t\geq 0}$ suit le procédé de Yule de paramètre λ, alors pour tous $t \geq 0$ et $n \in \mathbb{N}^*$, $\mathbb{P}(X_t = n) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}$.

Démonstration : Soit $t \geq 0$ et $n \in \mathbb{N}^*$.

- Si $n = 1$, alors, comme τ_1 suit la loi exponentielle de paramètre λ :

$$\mathbb{P}(X_t = n) = \mathbb{P}(\tau_1 > t) = e^{-\lambda t}.$$

- Supposons maintenant $n \geq 2$.
  Pour $k \geq 1$, τ_k suit la loi exponentielle de paramètre $k\lambda$, donc admet pour densité la fonction $x \longmapsto k\lambda e^{-k\lambda x}\mathbb{1}_{x\geq 0}$. Donc :

$$\begin{aligned}\mathbb{P}(X_t = n) &= \mathbb{P}(\tau_1 + \cdots + \tau_{n-1} \leq t \text{ et } \tau_1 + \cdots + \tau_n > t) \\ &= \int_{\mathbb{R}^n} \mathbb{1}_{t_1+\cdots+t_{n-1}\leq t} \cdot \mathbb{1}_{t_1+\cdots+t_n>t} \cdot \mathbb{1}_{t_1,\ldots,t_n\geq 0} \cdot \lambda e^{-\lambda t_1} dt_1 \cdots (n+1)\lambda e^{-n\lambda t_n} dt_n\end{aligned}$$

et avec le changement de variable $(v_1, v_2, \ldots, v_n) = (t_1, t_1 + t_2, \ldots, t_1 + \cdots + t_n)$,

$$\begin{aligned}
&= \int_{\mathbb{R}^n} \mathbb{1}_{0\leq v_1\leq\cdots\leq v_{n-1}\leq t\leq v_n} n!\ \lambda^n e^{-n\lambda v_n} e^{\lambda(v_1+\cdots+v_{n-1})} dv_1\cdots dv_n \\
&= \int_{\mathbb{R}^{n-1}} \mathbb{1}_{0\leq v_1\leq\cdots\leq v_{n-1}\leq t}\ n!\ \lambda^n \left(\int_t^{+\infty} e^{-n\lambda v_n} dv_n\right) e^{\lambda(v_1+\cdots+v_{n-1})} dv_1\cdots dv_{n-1} \\
&= e^{-n\lambda t} \int_{\mathbb{R}^{n-1}} \mathbb{1}_{0\leq v_1\leq\cdots\leq v_{n-1}\leq t}\ (n-1)!\ \lambda^{n-1} e^{\lambda(v_1+\cdots+v_{n-1})} dv_1\cdots dv_{n-1}.
\end{aligned}$$

Il suffit alors de vérifier que pour tous $k \in \mathbb{N}^*$ et $x \geq 0$ on a

$$I_k(x) := \int_{\mathbb{R}^k} \mathbb{1}_{0\leq v_1\leq\cdots\leq v_k\leq x}\ k!\ \lambda^k e^{\lambda(v_1+\cdots+v_k)} dv_1\cdots dv_k = (e^{\lambda x} - 1)^k$$

et on pourra alors conclure en prenant $k = n-1$ et $x = t$. Montrons donc ce résultat par récurrence sur k.

* $k = 1$: On a bien
$$I_1(x) = \int_0^x \lambda e^{\lambda v_1} dv_1 = e^{\lambda x} - 1.$$

* $k - 1 \to k$: Soit $k \geq 2$, supposons le résultat vrai pour $k-1$, alors

$$\begin{aligned}
I_k(x) &= \int_0^x \left(\int_{\mathbb{R}^{k-1}} \mathbb{1}_{0\leq v_1\leq\cdots\leq v_k} (k-1)!\ \lambda^{k-1} e^{\lambda(v_1+\cdots+v_{k-1})} dv_1\cdots dv_{k-1}\right) k\lambda e^{\lambda v_k} dv_k \\
&= \int_0^x (e^{\lambda v_k} - 1)^{k-1} k\lambda e^{\lambda v_k} dv_k \\
&= \Big[(e^{\lambda v_k} - 1)^k\Big]_0^x \\
&= (e^{\lambda x} - 1)^k - 0.
\end{aligned}$$

Ceci clôt la récurrence et donc la démonstration. ■

Yule utilise ce procédé de naissance linéaire pour décrire l'évolution des espèces et des genres. Cette évolution est régie par les deux points suivants :

1. Le nombre d'espèces dans un genre suit le procédé de Yule de paramètre λ.
2. Le nombre de genre croît de manière exponentielle de paramètre μ.

Proposition 7

Pour $t \geq 0$, on note N_t la variable aléatoire correspondant au nombre d'espèces d'un genre tiré uniformément parmi les genres existants à l'instant t. Alors N_t converge en loi quand $t \to +\infty$ vers une variable aléatoire $N^{(\rho)}$, dont la loi, appelée *distribution de Yule*, ne dépend que du paramètre $\rho = \lambda/\mu$ et vérifie :

$$\forall n \geq 1, \mathbb{P}(N^{(\rho)} = n) = \frac{\Gamma(1 + \frac{1}{\rho})}{\rho} \cdot \frac{\Gamma(n)}{\Gamma(n + 1 + \frac{1}{\rho})}$$

Démonstration :

- Pour $t \geq 0$, on note $G(t)$ le nombre de genres à l'instant t.
  D'après 1., $G(0) = 1$ et, d'après 2., $\forall t \geq 0, G'(t) = \mu G(t)$, donc $\forall t \geq 0, G(t) = e^{\mu t}$.

- Soit $t > 0$, on note T_t la variable aléatoire correspondant à l'âge du genre tiré uniformément parmi les $G(t)$ genres existants à l'instant t (le genre dont N_t est le nombre d'espèces).
  Soit $0 \leq s < t$, la probabilité que $T_t \in [o; s[$ est le rapport entre le nombre de genres nés entre $t-s$ et t et le nombre total de genres à l'instant t :

$$\mathbb{P}(T_t \in [o; s[) = \frac{G(t) - G(t-s)}{G(t)} = \frac{e^{\mu t} - e^{\mu(t-s)}}{e^{\mu t}} = 1 - e^{-\mu s} = \int_0^s \mu e^{-\mu u} du.$$

  D'autre part, on a $\mathbb{P}(T_t = t) = e^{-\mu t}$ et $\mathbb{P}(T_t < 0) = 0 = \mathbb{P}(T_t > t)$.
  Donc pour $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ mesurable,

$$\mathbb{E}[f(T_t)] = \int_0^t f(u)\mu e^{-\mu u} du + f(t)e^{-\mu t}.$$

- Soit $n \geq 1$, $\mathbb{P}(N_t = n) = \mathbb{E}\big[\mathbb{E}[\mathbb{1}_{N_t=n}|T_t]\big]$.
  Montrons que $\mathbb{E}[\mathbb{1}_{N_t=n}|T_t] = e^{-\lambda T_t}(1 - e^{-\lambda T_t})^{n-1}$. Soit $\varphi : \mathbb{R} \longrightarrow \mathbb{R}_+$ mesurable,

$$\mathbb{E}[\varphi(T_t)\mathbb{1}_{N_t=n}] = \sum_{k\geq 1} \int_{\mathbb{R}_+} \varphi(u)\mathbb{1}_{k=n} P_{(T_t,N_t)}(du, k) = \int_{\mathbb{R}_+} \varphi(u) P_{(T_t,N_t)}(du, n).$$

  Or pour $0 \leq u < t$,

$$\begin{aligned} P_{(T_t,N_t)}([u, u+du], n) &= \mathbb{P}(T_t \in [u, u+du] \text{ et } N_t = n) \\ &= \mathbb{P}(N_t = n | T_t \in [u, u+du])\mathbb{P}(T_t \in [u, u+du]) \\ &= e^{-\lambda u}(1 - e^{-\lambda u})^{n-1}\mu e^{-\mu u} du. \end{aligned}$$

  Et de même $P_{(T_t,N_t)}(t, n) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} e^{-\mu t}$ et pour $u < 0$ ou $u > t$, $P_{(T_t,N_t)}([u, u+du], n) = 0$, donc :

$$\begin{aligned} \mathbb{E}[\varphi(T_t)\mathbb{1}_{N_t=n}] &= \int_0^t \varphi(u)e^{-\lambda u}(1 - e^{-\lambda u})^{n-1}\mu e^{-\mu u} du + \varphi(t)e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} e^{-\mu t} \\ &= \mathbb{E}[\varphi(T_t)e^{-\lambda T_t}(1 - e^{-\lambda T_t})^{n-1}]. \end{aligned}$$

  Donc $\mathbb{E}[\mathbb{1}_{N_t=n}|T_t] = e^{-\lambda T_t}(1 - e^{-\lambda T_t})^{n-1}$ et on en déduit :

$$\begin{aligned} P(N_t = n) &= \mathbb{E}[e^{-\lambda T_t}(1 - e^{-\lambda T_t})^{n-1}] \\ &= \int_0^t e^{-\lambda u}(1 - e^{-\lambda u})^{n-1}\mu e^{-\mu u} du + e^{-\lambda t}(1 - e^{-\lambda t})^{n-1} e^{-\mu t} \\ &\underset{t\to+\infty}{\longrightarrow} \int_0^{+\infty} e^{-\lambda u}(1 - e^{-\lambda u})^{n-1}\mu e^{-\mu u} du. \end{aligned}$$

  Donc $N_t \xrightarrow{loi} N^{(\rho)}$ où $N^{(\rho)}$ a pour loi :

$$\begin{aligned} P(N^{(\rho)} = n) &= \int_0^{+\infty} e^{-(\lambda+\mu)u}(1 - e^{-\lambda u})^{n-1}\mu du \\ &= \int_0^1 v^{1+\frac{\mu}{\lambda}}(1 - v)^{n-1}\mu \frac{dv}{\lambda v} \quad \text{en posant } v = e^{-\lambda u} \\ &= \frac{1}{\rho} \cdot B(1 + \frac{1}{\rho}, n) \\ &= \frac{1}{\rho} \cdot \frac{\Gamma(1 + \frac{1}{\rho}) \cdot \Gamma(n)}{\Gamma(n + 1 + \frac{1}{\rho})}. \end{aligned}$$

■

Remarque : Yule décrit initialement l'évolution du nombre de genres G_t dans une famille (le taxon immédiatement au-dessus du genre dans la classification linnéenne) par un procédé de naissance linéaire de paramètre μ. Il fait seulement ensuite une approximation consistant à supposer qu'il y a un très grand nombre de familles pour approcher G_t par son espérance $e^{\mu t}$, ce qui justifie l'évolution exponentielle du nombre de genres imposée par 2. :

$$\mathbb{E}[G_t] = \sum_{n\geq 1} ne^{-\mu t}(1-e^{-\mu t})^{n-1} = e^{-\mu t} \cdot \frac{1}{\left(1-(1-e^{-\mu t})\right)^2} = e^{\mu t}.$$

Cette distribution de Yule permet de retrouver les deux résultat observés sur les données : la valeur 1 est la plus fréquente et la distribution est à queue lourde, c'est-à-dire que, si F est sa fonction de répartition, alors F n'est pas majorée et vérifie $\forall y > 0, F(x+y)/F(x) \xrightarrow{x\to+\infty} 1$. En effet on a, quand $k \to +\infty$,

$$\frac{\Gamma(k)}{\Gamma(k+1+\frac{1}{\rho})} \sim \frac{\sqrt{2\pi(k-1)}\left(\frac{k-1}{e}\right)^{k-1}}{\sqrt{2\pi(k+\frac{1}{\rho})}\left(\frac{k+\frac{1}{\rho}}{e}\right)^{k+\frac{1}{\rho}}} \sim \frac{e^{1+\frac{1}{\rho}}}{k^{1+\frac{1}{\rho}}} \cdot \frac{\left(1-\frac{1}{k}\right)^{k-1}}{\left(1-\frac{1}{k\rho}\right)^{k+\frac{1}{\rho}}} \sim \frac{e^{1+\frac{1}{\rho}}}{k^{1+\frac{1}{\rho}}} \cdot \frac{e^{-1}}{e^{\frac{1}{\rho}}} \sim \frac{1}{k^{1+\frac{1}{\rho}}}.$$

C'est le terme général positif d'une série convergente, donc, pour $n, m \geq 1$, par critère de Cauchy :

$$F(n+m) - F(n) = \frac{\Gamma(1+\frac{1}{\rho})}{\rho} \cdot \sum_{k=n+1}^{n+m} \frac{\Gamma(k)}{\Gamma(k+1+\frac{1}{\rho})} \underset{n\to+\infty}{\longrightarrow} 0.$$

Donc $F(m+n)/F(n) \underset{n\to+\infty}{\longrightarrow} 1$ car $F(n) \underset{n\to+\infty}{\longrightarrow} 1$.

Au delà de ces deux propriétés, Yule compara également cette distribution aux données quantitatives de l'époque et il obtint des résultats très satisfaisant, comme, par exemple, dans le cas des serpents pour le tableau suivant ([2], Table 1).

Number species in genus	**Number of genera**	
	Observed	**Calculated**
1	131	130.9
2	35	47.2
3	28	25.2
4	17	16.0
5	16	11.2
6	9	8.3
7	8	6.5
8	8	5.2
9 to 11	13	11.1
12 to 14	3	7.2
15 to 20	7	8.8
21 to 34	14	9.2
35 upward	4	6.2

2.2 Le modèle de Yule ou modèle ERM

Dans le cadre des arbres phylogénétiques, un genre se définirait comme le plus petit-clade au-dessus de l'espèce. Alors il n'y a pas de genres de taille supérieure ou égale à 3 car si un clade a au moins 3 espèces, alors il a forcément un sous-clade d'au moins 2 espèces et il ne peut donc pas être le genre de ces 2 espèces. Donc, si on essaye de définir des genres qui soient des clades, alors ces genres seront tous de taille 1 ou 2. Les genres ne semblent donc plus être un critère pertinent pour comparer les formes des arbres phylogénétiques.

Sans parler de genres, on peut se servir de nouveau du procédé de Yule pour définir une suite de distributions sur les cladogrammes à partir de (b), appelée *modèle de Yule*. On prend $(X_t)_{t\geq 0}$ une population d'espèces qui suit le procédé de Yule de paramètre 1, et on arrête le processus dès qu'il y a n espèces : on obtient alors un cladogramme aléatoire à n feuilles en enlevant l'échelle temporelle et en attribuant aléatoirement les étiquettes $\{1,\dots,n\}$ aux espèces de manière uniforme. On note alors T_n^{Y} sa distribution.

Dans le but d'obtenir une famille à un paramètre, on pourrait essayer de modifier le paramètre du procédé de Yule dans le modèle ci-dessus. On va cependant considérer une classe de modèles encore plus grande, qui est celle des modèles que l'on peut décrire par le procédé suivant. A l'instant $t = 0$, il y a une espèce. Au fil du temps, il peut y avoir un évènement qui est soit une extinction soit une spéciation, c'est-à-dire, soit une espèce A s'éteint, soit une espèce B se sépare en deux espèces B et B'. Le temps entre un instant t et le prochain évènement et la probabilité que cet évènement soit une extinction ou une spéciation peut dépendre arbitrairement du nombre d'espèces qu'il y a eu dans le passé. Mais si le prochain évènement est une extinction alors chaque espèce a autant de chance d'être celle qui s'éteint, et si c'est une spéciation, alors chaque espèce a autant de chance d'être celle qui spécie. Pour $n \geq 1$, on arrête ce procédé au bout d'un nombre K_n d'évènements fixé, et alors on note T_n la distribution du cladogramme ainsi obtenu après avoir attribué les étiquettes $\{1,\dots,n\}$ aux espèces de manière uniforme.

Pour $k \geq 1$, on note τ_k l'instant du $k^{\text{ème}}$ événement (spéciation ou extinction) qui a lieu, et on fixe $\tau_0 = 0$: on a donc $0 = \tau_0 < \tau_1 < \cdots < \tau_k < \cdots$. On note alors X_k le cladogramme aléatoire obtenu à partir des espèces existante entre les instants τ_{k-1} et τ_k : ainsi X_k a au plus k feuilles. Par définition, pour $C \in \mathcal{C}_n$, la mesure de C sous ce modèle est $T_n(C) = \mathbb{P}(X_{K_n} = C \mid X_{K_n} \in \mathcal{C}_n)$.

Mais l'arbitraire laissé dans la définition de cette classe de modèle ne permet pas de construire une famille de distributions à un paramètre, car tous ces modèles induisent la même distribution que le modèle ERM introduit ci-dessous (proposition 8).

Le *modèle ERM* (pour *Equal Rate Markov*), qui correspond en fait à la description rétrograde des modèles précédents, est défini par le procédé suivant : on part de n espèces que l'on voit comme n lignées de descendance, on tire uniformément une paire parmi $n(n-1)/2$ paires de lignées et on joint ces deux lignées pour obtenir $n-1$ lignées, puis on réitère ce processus jusqu'à n'avoir plus qu'une seule lignée et on obtient alors un cladogramme à n feuilles. On note T_n^{ERM} la distribution de ce cladogramme aléatoire.

Proposition 8

Toutes les suites $(T_n)_{n\geq 1}$ de distributions obtenues à partir des modèles de cette classe sont identiques à la suite $(T_n^{\mathrm{ERM}})_{n\geq 1}$ et donc en particulier à la suite $(T_n^{\mathrm{Y}})_{n\geq 1}$.

Nous montrerons ce résultat en 2.3.

Le fait que tous ces modèles soient identiques signifie que la forme des arbres phylogénétiques ne dépend pas des taux globaux d'extinction et de spéciation si l'on suppose qu'à chaque évènement toutes les espèces ont autant de chance d'être touchées. Un changement de ces taux modifierait l'échelle temporelle des arbres, mais cette échelle disparaît dans le cladogramme.

2.3 Propriétés du modèle

Nous nous intéressons dans cette partie aux propriétés mathématiques du modèle ERM. La proposition suivante montre que même si ce modèle est construit à partir de (b), il vérifie les propriétés (i) et (ii).

Proposition 9

La suite de distributions $(T_n^{\mathrm{ERM}})_{n\geq 1}$ sur les $\mathcal{C}_n$ vérifie les propriétés (i) et (ii).

Démonstration :

(i) La propriété (i) est évidente car toutes les paires sont équiprobables.

(ii) Soit $n \geq 1$, on note d_n le nombre de procédés possibles pour obtenir un cladogramme à n feuilles (c'est-à-dire le nombre de suites de $n-1$ paires de lignées que l'on peut choisir). Pour $C \in \mathcal{C}_n$, on note $d(C)$ le nombre de procédés qui permettent de construire C. Alors $T_n^{\mathrm{ERM}}(C) = d(C)/d_n$. De même, on note $d(\mathcal{B})$ pour $\mathcal{B}$ une partie de $\mathcal{C}_n$.
Soient $1 \leq k < n$ et $C \in \mathcal{C}_k$. On veut montrer que $T_n^{\mathrm{ERM}}(\varphi_{n,k}^{-1}(C))/T_n^{\mathrm{ERM}}(\mathcal{A}_{n,k}) = T_k(C)$. Il suffit de montrer que $d(\varphi_{n,k}^{-1}(C))/d(\mathcal{A}_{n,k}) = d(C)/d_k$.
Se donner un procédé construisant un cladogramme de $\mathcal{A}_{n,k}$ revient à :

- choisir un procédé pour construire un arbre $C_1 \in \mathcal{C}_k$: d_k possibilités ;
- choisir un procédé pour construire un arbre $C_2 \in \mathcal{C}_{n-k}$ (pour lequel on remplace les étiquettes $\{1,\dots,n-k\}$ des espèces par $\{k+1,\dots,n\}$) : d_{n-k} possibilités ;
- choisir une manière de fabriquer un procédé pour construire, en conservant l'ordre des procédés de C_1 et C_2, un arbre C' à n feuilles qui contient C_2 comme sous-arbre et pour lequel C_1 est l'arbre obtenu en supprimant le sous-arbre C_2 à C' (et l'arête au-dessus de sa racine) : x possibilités (ne dépend pas de C_1 et C_2 ni des procédés, mais uniquement de leur longueur, qui est fixe).

Donc $d(\mathcal{A}_{n,k}) = d_k \cdot d_{n-k} \cdot x$.
On peut facilement expliciter une expression de x, mais ici ce n'est pas nécessaire, car x apparaît également dans le calcul de $d(\varphi_{n,k}^{-1}(C))$.

En effet, se donner un procédé construisant un cladogramme de $\varphi_{n,k}^{-1}(C)$ revient à :

- choisir un procédé pour construire C : $d(C)$ possibilités ;
- choisir un procédé pour construire un arbre $C_2 \in \mathcal{C}_{n-k}$ (pour lequel on remplace les étiquettes $\{1,\dots,n-k\}$ des espèces par $\{k+1,\dots,n\}$) : d_{n-k} possibilités ;

∗ choisir une manière de fabriquer un procédé pour construire, en conservant l'ordre des procédés de C et C_2, un arbre C' à n feuilles qui contient C_2 comme sous-arbre et pour lequel C est l'arbre obtenu en supprimant le sous-arbre C_2 : x possibilités.

Donc $d(\varphi_{n,k}^{-1}(C)) = d(C) \cdot d_{n-k} \cdot x$ et donc on a bien $d(\varphi_{n,k}^{-1}(C))/d(\mathcal{A}_{n,k}) = d(C)/d_k$. ■

Le lemme qui suit permet de calculer simplement la probabilité d'un cladogramme sous le modèle ERM et sera utilisé par la suite (en 4.4).

Lemme 10

Soient $n \geq 2$ et $C \in \mathcal{C}_n$. Si k et $n-k$ sont les tailles des deux sous-clades de la racine de C et si $C_k \in \mathcal{C}_k$ et $C_{n-k} \in \mathcal{C}_{n-k}$ sont ces deux sous-clades, alors

$$T_n^{\mathrm{ERM}}(C) = \frac{2}{(n-1)\binom{n}{k}} \cdot T_k^{\mathrm{ERM}}(C_k) \cdot T_{n-k}^{\mathrm{ERM}}(C_{n-k}).$$

Cette relation de récurrence caractérise $(T_n^{\mathrm{ERM}})_{n\geq 1}$ car si pour N un nœud interne de C, on note m_N la taille du clade associé à N et k_N la taille de son plus petit sous-clade, alors

$$T_n^{\mathrm{ERM}}(C) = \prod_{N \text{ nœud de } C} \frac{2}{(m_N-1)\binom{m_N}{i_N}}.$$

Définis ainsi, C_k et C_{n-k} ne portent pas les étiquettes $\{1,\dots,k\}$ et $\{1,\dots,n-k\}$, donc n'appartiennent pas vraiment à $\mathcal{C}_k$ et $\mathcal{C}_{n-k}$, mais leurs formes sont bien définies et, comme $(T_n^{\mathrm{ERM}})_{n\geq 1}$ vérifie (i), quelle que soit la manière dont on modifie leurs étiquettes pour qu'ils appartiennent à $\mathcal{C}_k$ et $\mathcal{C}_{n-k}$, l'égalité ci-dessus reste vraie.

Démonstration :

- Soit $n \geq 2$, d_n est le nombre de suites de $n-1$ paires de lignées que l'on peut choisir, donc on a :

$$d_n = \binom{n}{2} \cdot \binom{n-1}{2} \cdots \binom{2}{2} = \frac{n(n-1)}{2} \cdot \frac{(n-1)(n-2)}{2} \cdots \frac{2\cdot 1}{2} = \frac{n}{2^{n-1}} \cdot \big((n-1)!\big)^2.$$

  Soit $C \in \mathcal{C}_n$, on note k et $n-k$ les tailles des deux sous-clades de la racine de C et $C_k \in \mathcal{C}_k$ et $C_{n-k} \in \mathcal{C}_{n-k}$ ces deux sous-clades. Choisir un procédé pour construire C revient à se donner :

  ∗ un procédé pour construire C_k (qui a donc $k-1$ étapes) : $d(C_k)$ possibilités ;
  ∗ un procédé pour construire C_{n-k} (qui a donc $n-k-1$ étapes) : $d(C_{n-k})$ possibilités ;
  ∗ une partie X à $k-1$ éléments de $[\![1, n-2]\!]$ (le procédé pour construire C consiste alors, pour i allant de 1 à $n-2$, à effectuer une étape du procédé de C_k si $i \in X$ ou du procédé de C_{n-k} sinon (les deux procédés étant ainsi effectués dans l'ordre en parallèle) et à terminer par l'étape qui relie C_k et C_{n-k}) : $\binom{n-2}{k-1}$ possibilités.

On a donc $d(C) = \binom{n-2}{k-1} \cdot d(C_k) \cdot d(C_{n-k})$ et on en déduit :

$$\begin{aligned} T_n^{\mathrm{ERM}}(C) &= \frac{\binom{n-2}{k-1} \cdot d(C_k) \cdot d(C_{n-k})}{d_n} \\ &= \frac{\binom{n-2}{k-1} \cdot d_k \cdot d_{n-k}}{d_n} \cdot T_k^{\mathrm{ERM}}(C_k) \cdot T_{n-k}^{\mathrm{ERM}}(C_{n-k}) \\ &= \frac{(n-2)! \cdot 2^{n-1} \cdot k!(k-1)! \cdot (n-k)!(n-k-1)!}{(n-k-1)!(k-1)! \cdot n!(n-1)! \cdot 2^{k-1} \cdot 2^{n-k-1}} \cdot T_k^{\mathrm{ERM}}(C_k) \cdot T_{n-k}^{\mathrm{ERM}}(C_{n-k}) \\ &= \frac{2}{(n-1)\binom{n}{k}} \cdot T_k^{\mathrm{ERM}}(C_k) \cdot T_{n-k}^{\mathrm{ERM}}(C_{n-k}). \end{aligned}$$

- Alors par une récurrence forte sur n, on obtient facilement l'expression de $T_n^{\mathrm{ERM}}(C)$, en se ramenant à des cladogrammes à une feuille qui ont une probabilité 1. ■

On va maintenant montrer la proposition 8. On va pour cela montrer les deux lemmes suivants.

Lemme 11

Le modèle de Yule et le modèle ERM sont identiques : $(T_n^{\mathrm{ERM}})_{n\geq 1} = (T_n^{\mathrm{Y}})_{n\geq 1}$.

Démonstration : Comme dans la définition de la classe de modèle en 2.2, on définit $(X_k)_{k\geq 1}$ la suite de cladogrammes aléatoires obtenue par le procédé du modèle de Yule. On a ici $X_n \in \mathcal{C}_n$ parce qu'il n'y a que des spéciations et pour $C \in \mathcal{C}_n$, $T_n^{\mathrm{Y}}(C) = \mathbb{P}(X_n = C)$.

- La suite $(T_n^{\mathrm{Y}})_{n\geq 1}$ vérifie la propriété (i) par définition : une fois la forme fixée par le processus, la répartition des étiquettes est uniforme sur l'ensemble des possibilités. Le poids d'un cladogramme sous T_n^{Y} ne dépend donc que de sa forme.
- Soit $1 \leq i \leq n/2$. On pose $\mathcal{I}_i^n$ l'ensemble des cladogrammes à n feuilles dont le plus-petit sous-clade de la racine a i feuilles.
  Montrons par récurrence sur $n \geq 1$ que

$$T_n^{\mathrm{Y}}(\mathcal{I}_i^n) = \begin{cases} \frac{2}{n-1} & \text{si } i \neq n/2 \\ \frac{1}{n-1} & \text{si } i = n/2 \end{cases}.$$

  C'est évident pour $n = 1$.
  Soit $n \geq 2$, supposons le résultat vrai au rang $n-1$. Supposons $i \neq n/2$ et $i \neq (n-1)/2$.
  Si $X_n \in \mathcal{I}_i^n$, alors $X_{n-1} \in \mathcal{I}_i^{n-1}$ ou $X_{n-1} \in \mathcal{I}_{i-1}^{n-1}$ et ces évènements sont disjoints, donc :

$$\begin{aligned} \mathbb{P}(X_n \in \mathcal{I}_i^n) = \mathbb{P}(X_n \in \mathcal{I}_i^n \mid X_{n-1} \in \mathcal{I}_i^{n-1})\ \mathbb{P}(X_{n-1} \in \mathcal{I}_i^{n-1}) \\ + \mathbb{P}(X_n \in \mathcal{I}_i^n \mid X_{n-1} \in \mathcal{I}_{i-1}^{n-1})\ \mathbb{P}(X_{n-1} \in \mathcal{I}_{i-1}^{n-1}). \end{aligned}$$

  Par hypothèse de récurrence, $\mathbb{P}(X_{n-1} \in \mathcal{I}_i^{n-1}) = 2/(n-2)$ et $\mathbb{P}(X_{n-1} \in \mathcal{I}_{i-1}^{n-1}) = 2/(n-2)$ car $i \neq (n-1)/2$.
  D'autre part, $\mathbb{P}(X_n \in \mathcal{I}_i^n \mid X_{n-1} \in \mathcal{I}_i^{n-1})$ est la probabilité que ce soit une espèce du clade de X_{n-1} à $n-i-1$ éléments qui spécie et $\mathbb{P}(X_n \in \mathcal{I}_i^n \mid X_{n-1} \in \mathcal{I}_{i-1}^{n-1})$ est la probabilité que ce soit une espèce du clade de X_{n-1} à $i-1$ éléments qui spécie.

Donc on obtient :

$$T_n^{\mathrm{Y}}(\mathcal{I}_i^n) = \mathbb{P}(X_n \in \mathcal{I}_i^n) = \frac{n-i-1}{n-1} \cdot \frac{2}{n-2} + \frac{i-1}{n-1} \cdot \frac{2}{n-2} = \frac{2}{n-1}.$$

Comme $\mathcal{C}_n$ est l'union disjointe des $\mathcal{I}_i^n$ pour $1 \leq i \leq \lfloor n/2 \rfloor$, on en déduit la résultat pour la dernière valeur de i en utilisant que T_n^{Y} est de masse totale 1 : si n est impair, pour $i = (n-1)/2$, $T_n^{\mathrm{Y}}(\mathcal{I}_i^n) = 2/(n-1)$ et si n est impair, pour $i = n/2$, $T_n^{\mathrm{Y}}(\mathcal{I}_i^n) = 1/(n-1)$.

- Suite à la première spéciation à l'instant τ_1, X_2 est l'unique cladogramme à deux espèces. Supposons qu'à cet instant on fixe arbitrairement l'orientation gauche droite de la racine (avec probabilité $1/2$ pour chaque possibilité). Alors chacune des deux espèces filles se reproduisent selon des procédés de Yule indépendants : les deux cladogrammes X_n^g et X_n^d associés aux sous-clades gauches et droites de la racine sont indépendants et suivent chacun le modèle de Yule (si on réindice ses procédés par les instants où ils gagnent une feuille, en partant de 1). Le fait que ces sous-cladogrammes n'aient pas pour étiquettes un ensemble $[\![1, k]\!]$ mais un sous-ensemble de $[\![1, n]\!]$, n'empêche pas de définir leur poids sous T_k^{Y} car le modèle de Yule vérifie la propriété (i).
  On note $q_n(k)$ la probabilité que X_n^g ait k feuilles.
  Si $k \neq n/2$, on note $i = \min(k, n-k)$, alors $q_n(k) = T_n^{\mathrm{Y}}(\mathcal{I}_i^n)/2 = 1/(n-1)$ car la clade de gauche a une chance sur deux d'être la plus petite.
  Si $k = n/2$, alors $q_n(k) = T_n^{\mathrm{Y}}(\mathcal{I}_{k/2}^n) = 1/(n-1)$.
  Comme les deux sous-clades se comportent récursivement selon le même modèle, on vient de vérifier que $(T_n^{\mathrm{Y}})_{n\geq 1}$ était le modèle de branchements de Markov associé à $(q_n)_{n\geq 2}$ (cf définition 19) et on peut donc appliquer le lemme 20 qui montre, dans un cadre plus général, que si $n \geq 2$ et $C \in \mathcal{C}_n$, si k et $n-k$ sont les tailles des deux sous-clades de la racine de C et si $C_k \in \mathcal{C}_k$ et $C_{n-k} \in \mathcal{C}_{n-k}$ sont ces deux sous-clades, alors

$$T_n^{\mathrm{Y}}(C) = \frac{2 \cdot q_n(k)}{\binom{n}{k}} \cdot T_k^{\mathrm{Y}}(C_k) \cdot T_{n-k}^{\mathrm{Y}}(C_{n-k}) = \frac{2}{(n-1)\binom{n}{k}} \cdot T_k^{\mathrm{Y}}(C_k) \cdot T_{n-k}^{\mathrm{Y}}(C_{n-k}).$$

  Or cette relation de récurrence caractérise $(T_n^{\mathrm{ERM}})_{n\geq 1}$ d'après le lemme 10 donc $(T_n^{\mathrm{Y}})_{n\geq 1} = (T_n^{\mathrm{ERM}})_{n\geq 1}$. ■

Lemme 12

Soient $n \geq 2$, X une variable aléatoire à valeurs dans $\mathcal{C}_n$ de loi T_n^{ERM} et I une variable aléatoire uniforme sur $[\![1, n]\!]$ indépendante de X.

- Soit Y le cladogramme aléatoire à $n-1$ feuilles obtenu en supprimant l'espèce I de X, puis en resdistribuant de manière uniforme les étiquettes $\{1, \ldots, n-1\}$ sur les feuilles du cladogramme ainsi obtenu. Alors la loi de Y est T_{n-1}^{ERM}.
- Soit Z le cladogramme aléatoire à $n+1$ feuilles obtenu en spéciant l'espèce I de X (la feuille I est remplacée par un clade de 2 espèces), puis en resdistribuant de manière uniforme les étiquettes $\{1, \ldots, n+1\}$ sur les feuilles du cladogramme ainsi obtenu. Alors la loi de Z est T_{n+1}^{ERM}.

Démonstration :

- Soit $C \in \mathcal{C}_{n-1}$. Comme le modèle ERM vérifie la propriété (ii), on a

$$T_{n-1}^{\mathrm{ERM}}(C) = \frac{T_n^{\mathrm{ERM}}(\varphi_{n,n-1}^{-1}(C))}{T_n^{\mathrm{ERM}}(\mathcal{A}_{n,n-1})} = T_n^{\mathrm{ERM}}(\varphi_{n,n-1}^{-1}(C))$$

car $\mathcal{A}_{n,n-1}$ est l'ensemble des $C' \in \mathcal{C}_n$ tels que $\{n\}$ est un clade dans C', donc $\mathcal{A}_{n,n-1} = \mathcal{C}_n$. On a donc $\mathbb{P}(\varphi_{n,n-1}(X) = C) = T_{n-1}^{\mathrm{ERM}}(C)$.
On rappelle que pour $C' \in \mathcal{C}_n$ et $\sigma \in \mathcal{S}_n$, C'_σ est le cladogramme obtenu par permutation σ des espèces. Comme le modèle ERM vérifie la propriété (i), on a $X \overset{\text{loi}}{=} X_\sigma$. Pour $i \in [\![1, n]\!]$, on note $\sigma_i = (i\ n)$. Alors l'arbre obtenu en supprimant l'espèce i de X est de la forme de $\varphi_{n,n-1}(X_{\sigma_i})$. Si $I = i$, Y est égal à $\varphi_{n,n-1}(X_{\sigma_i})$ à une permutation des espèces près qui est tirée uniformément dans $\mathcal{S}_{n-1}$, donc

$$\begin{aligned}
\mathbb{P}(Y = C) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(Y = C \mid I = i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(n-1)!} \sum_{\sigma \in \mathcal{S}_{n-1}} \mathbb{P}(\big(\varphi_{n,n-1}(X_{\sigma_i})\big)_\sigma = C) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(n-1)!} \sum_{\sigma \in \mathcal{S}_{n-1}} \mathbb{P}(\varphi_{n,n-1}(X_{\tilde{\sigma}\circ\sigma_i}) = C) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(n-1)!} \sum_{\sigma \in \mathcal{S}_{n-1}} T_{n-1}^{\mathrm{ERM}}(C) \ \text{ car } X \overset{\text{loi}}{=} X_{\tilde{\sigma}\circ\sigma_i} \\
&= T_{n-1}^{\mathrm{ERM}}(C)
\end{aligned}$$

où $\tilde{\sigma}$ est la permutation de $[\![1, n]\!]$ dont la restriction à $[\![1, n-1]\!]$ est σ.
On a donc montré que Y a pour loi T_{n-1}^{ERM}.

- Cette propriété est une conséquence directe du lemme 11. En effet, on sait que X a la même loi que X_n le cladogramme du modèle de Yule obtenu après $n-1$ spéciations. Or X_{n+1} est par définition le cladogramme obtenu à partir de X_n après une spéciation supplémentaire, touchant une espèces tirée uniformément parmi les n espèces de X_n. Donc Z a la loi de X_{n+1}, c'est-à-dire $T_{n+1}^{\mathrm{Y}} = T_{n+1}^{\mathrm{ERM}}$. ■

Démonstration de la proposition 8 :

On considère un modèle appartenant à la classe de modèles définie en 2.2, on note $(T_n)_{n\geq 1}$ la suite de distributions associée et X_k le cladogramme aléatoire obtenu après le $(k-1)^{\text{ème}}$ événement. Par définition, pour $C \in \mathcal{C}_n$, la mesure de C sous ce modèle est $T_n(C) = \mathbb{P}(X_{K_n} = C \mid X_{K_n} \in \mathcal{C}_n)$. On va montrer plus généralement que pour tous $n, k \in \mathbb{N}^*$, $\mathbb{P}(X_k = C \text{ et } X_k \in \mathcal{C}_n) = T_n^{\mathrm{ERM}}(C) \cdot \mathbb{P}(X_k \in \mathcal{C}_n)$ et on aura donc $(T_n)_{n\geq 1} = (T_n^{\mathrm{ERM}})_{n\geq 1}$.

Pour cela on procède par récurrence sur $k \geq 1$ avec pour hypothèse de récurrence $\mathcal{H}_k$: « Pour tous $n \geq 1$, $f : \mathbb{N}^{k-1} \longrightarrow \mathbb{R}_+$ et $F : \mathcal{C}_n \longrightarrow \mathbb{R}_+$ mesurables,

$$\mathbb{E}\big[f(\#X_1, \ldots, \#X_{k-1})F(X_k)\mathbb{1}_{\#X_k=n}\big] = \mathbb{E}_n^{\mathrm{ERM}}[F] \cdot \mathbb{E}\big[f(\#X_1, \ldots, \#X_{k-1})\mathbb{1}_{\#X_k=n}\big] \text{ »}$$

où $\#X_i$ est le nombre de feuilles de X_i et $\mathbb{E}_n^{\mathrm{ERM}}[F]$ est l'espérance de F vue comme une variable aléatoire définie sur $(\mathcal{C}_n, T_n^{\mathrm{ERM}})$.

- Pour $k = 1$, X_1 est l'unique cladogramme à 1 feuilles. On a donc directement $\mathcal{H}_1$ vraie.
- Soit $k \geq 1$, supposons $\mathcal{H}_k$.
  Soient $n \geq 1$, $f : \mathbb{N}^{k-1} \longrightarrow \mathbb{R}_+$ et $F : \mathcal{C}_n \longrightarrow \mathbb{R}_+$ mesurables.
  X_{k+1} est le cladogramme obtenu à partir de X_k suite au $k^{\text{ème}}$ évènement. On note $S_k = \{\text{le } k^{\text{ème}} \text{ évènement est une spéciation}\}$ et $E_k = \{\text{le } k^{\text{ème}} \text{ évènement est une extinction}\}$.

$$\begin{aligned}\mathbb{E}\big[f(\#X_1,\dots,\#X_k)F(X_{k+1})\mathbb{1}_{\#X_{k+1}=n}\big] =& \mathbb{E}\big[f(\#X_1,\dots,\#X_k)F(X_{k+1})\mathbb{1}_{\#X_{k+1}=n}\mathbb{1}_{S_k}\big] \\ &+ \mathbb{E}\big[f(\#X_1,\dots,\#X_k)F(X_{k+1})\mathbb{1}_{\#X_{k+1}=n}\mathbb{1}_{E_k}\big]\end{aligned}$$

  Notons A et B les deux termes de droite dans l'égalité ci-dessus.
  Sous les évènements S_k et $\{\#X_{k+1} = n\}$, on a $\#X_k = n-1$ et X_{k+1} est obtenu à partir de X_k en spéciant l'espèce I de X_k, où I est une variable aléatoire uniforme sur $[\![1, n-1]\!]$ indépendante de $X_1,\dots,X_k$, et en redistribuant les étiquettes uniformément, indépendamment de $X_1,\dots,X_k$ et I. Alors on a

$$A = \mathbb{E}\big[f(\#X_1,\dots,\#X_k)\tilde{F}(X_k)\mathbb{1}_{S_k}\mathbb{1}_{\#X_{k+1}=n}\big]$$

  où $\tilde{F} : \mathcal{C}_{n-1} \longrightarrow \mathbb{R}_+$ est définie, pour $C \in \mathcal{C}_{n-1}$ et $\psi_i(C)$ le cladogramme obtenu en remplaçant l'espèce i de C par le clade $\{i, n\}$, par

$$\tilde{F}(C) = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n} F\big((\psi_i(C))_\sigma\big)$$

  car on peut intégrer dans l'espérance par rapport à I et au tirage des étiquettes, qui sont indépendants entre eux et de $X_1,\dots,X_k$.
  Comme sous S_k on a $\{\#X_{k+1} = n\} = \{\#X_k = n-1\}$, on obtient

$$\begin{aligned}A &= \mathbb{E}\big[f(\#X_1,\dots,\#X_{k-1},n-1)\mathbb{1}_{S_k}\tilde{F}(X_k)\mathbb{1}_{\#X_k=n-1}\big] \\ &= \mathbb{E}^{\text{ERM}}_{n-1}[\tilde{F}]\cdot\mathbb{E}\big[f(\#X_1,\dots,\#X_{k-1},n-1)\mathbb{1}_{S_k}\mathbb{1}_{\#X_k=n-1}\big]\end{aligned}$$

  d'après $\mathcal{H}_k$ car $\mathbb{1}_{S_k}$ est une fonction du nombre d'espèces qu'il y a eu avant τ_k, c'est-à-dire une fonction de $\#X_1,\dots,\#X_k$ qui sous $\{\#X_k = n-1\}$ devient une fonction de $\#X_1,\dots,\#X_{k-1}$.
  Par définition, $\mathbb{E}^{\text{ERM}}_{n-1}[\tilde{F}] = \mathbb{E}[\tilde{F}(X)]$ où X est une variable aléatoire à valeurs dans $\mathcal{C}_{n-1}$ de loi T^{ERM}_{n-1}. Par construction de $\tilde{F}$, $\mathbb{E}[\tilde{F}(X)] = \mathbb{E}[F(Z)]$ où Z est la variable aléatoire à valeurs dans $\mathcal{C}_n$ définie dans le $2^{\text{ème}}$ point du lemme 12. D'après ce lemme, on a $\mathbb{E}[F(Z)] = \mathbb{E}^{\text{ERM}}_n[F]$.
  En utilisant de nouveau que sous S_k, on a $\{\#X_{k+1} = n\} = \{\#X_k = n-1\}$, on obtient

$$A = \mathbb{E}^{\text{ERM}}_n[F]\cdot\mathbb{E}\big[f(\#X_1,\dots,\#X_k)\mathbb{1}_{S_k}\mathbb{1}_{\#X_{k+1}=n}\big].$$

  On raisonne de manière analogue pour calculer B. Sous les évènements E_k et $\{\#X_{k+1} = n\}$, on a $\#X_k = n+1$ et X_{k+1} est obtenu à partir de X_k en supprimant l'espèce I de X_k, où I est une variable aléatoire uniforme sur $[\![1, n+1]\!]$ indépendante de $X_1,\dots,X_k$, et en redistribuant les étiquettes uniformément, indépendamment de $X_1,\dots,X_k$ et I. On a alors

$$B = \mathbb{E}\big[f(\#X_1,\dots,\#X_k)\hat{F}(X_k)\mathbb{1}_{E_k}\mathbb{1}_{\#X_{k+1}=n}\big]$$

où $\hat{F} : \mathcal{C}_{n+1} \longrightarrow \mathbb{R}_+$ est définie, pour $C \in \mathcal{C}_{n-1}$ et avec les notations de la démonstration du lemme 12, par

$$\hat{F}(C) = \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} F\big(\varphi_{n,n-1}(X_{\tilde{\sigma} \circ \sigma_i})\big).$$

De même, en utilisant $\mathcal{H}_k$ et le 1er point du lemme 12, on obtient

$$\begin{aligned} B &= \mathbb{E}_{n+1}^{\mathrm{ERM}}[\hat{F}] \cdot \mathbb{E}\big[f(\#X_1, \ldots, \#X_k) \mathbb{1}_{E_k} \mathbb{1}_{\#X_k = n-1}\big] \\ &= \mathbb{E}_n^{\mathrm{ERM}}[F] \cdot \mathbb{E}\big[f(\#X_1, \ldots, \#X_k) \mathbb{1}_{E_k} \mathbb{1}_{\#X_{k+1} = n}\big]. \end{aligned}$$

Ceci montre $\mathcal{H}_{k+1}$ car, comme $\mathbb{1}_{S_k} + \mathbb{1}_{E_k} = 1$,

$$\begin{aligned} \mathbb{E}\big[f(\#X_1, \ldots, \#X_k) F(X_{k+1}) \mathbb{1}_{\#X_{k+1} = n}\big] &= A + B \\ &= \mathbb{E}_n^{\mathrm{ERM}}[F] \cdot \mathbb{E}\big[f(\#X_1, \ldots, \#X_k) \mathbb{1}_{\#X_{k+1} = n}\big]. \end{aligned}$$

Ceci conclut la récurrence. Soient $n, k \in \mathbb{N}^*$ et $C \in \mathcal{C}_n$, on peut appliquer le résultat que l'on vient de montrer pour avoir

$$\begin{aligned} \mathbb{P}(X_k = C \text{ et } X_k \in \mathcal{C}_n) &= \mathbb{E}[\mathbb{1}_{X_k = C} \cdot \mathbb{1}_{\#X_k = n}] \\ &= \mathbb{E}_n^{\mathrm{ERM}}[F] \cdot \mathbb{E}[\mathbb{1}_{\#X_k = n}] \\ &= T_n^{\mathrm{ERM}}(C) \cdot \mathbb{P}(X_k \in \mathcal{C}_n), \end{aligned}$$

qui est ce que l'on voulait montrer. ■

Tous les modèles définis en 2.2 étant identiques, on utilisera dans la suite uniquement le nom *modèle ERM* pour les désigner tous.

3 Autres distributions particulières

La tentative d'obtenir une famille de distributions à un paramètre à partir de (b) ayant échoué (puisque qu'on ne récupère ainsi qu'un seul modèle), on peut essayer de construire d'autres modèles à partir des propriétés (i) et (ii).

3.1 Le modèle PDA

Définition 13

Le *modèle uniforme* ou *modèle PDA* (pour *Proportional to Different Arrangements*) a pour suite de distributions $(T_n^{\mathrm{PDA}})_{n \geq 1}$ où T_n^{PDA} est la probabilité uniforme sur $\mathcal{C}_n$. Pour $n \geq 1$, on pose $c_n = \#\mathcal{C}_n$ et ainsi $\forall C \in \mathcal{C}_n, T_n^{\mathrm{PDA}}(C) = 1/c_n$.

Lemme 14

Pour $n \geq 2$, $c_n = (2n-3)!! = 1 \cdot 3 \cdots (2n-5) \cdot (2n-3)$.

Démonstration : Un arbre binaire à n feuilles a $2n-1$ nœuds et donc $2n-2$ arêtes. On rajoute ici une arête au-dessus de la racine (il a alors $2n-1$ arêtes). Alors, pour $n \geq 1$, se donner un cladogramme à $n+1$ feuilles revient à :

- choisir un cladogramme à n feuilles (qui contient donc les espèces étiquetées de 1 à n) : c_n possibilités ;
- choisir une arête de ce cladogramme (on sépare alors cette arête en deux par un nœud, dont le deuxième fils est l'espèce étiquetée par $n+1$) : $2n-1$ possibilités.

Donc $c_{n+1} = (2n-1)c_n$ et on en déduit le résultat car $c_1 = 1$. ■

Proposition 15

La suite de distributions $(T_n^{\mathrm{PDA}})_{n\geq 1}$ sur les $\mathcal{C}_n$ du modèle PDA vérifie les propriétés (i) et (ii).

Démonstration :

(i) Pour tous $n \geq 1$, $C \in \mathcal{C}_n$ et $\sigma \in \mathcal{S}_n$, on a $T_n^{\mathrm{PDA}}(C) = 1/c_n = T_n^{\mathrm{PDA}}(C_\sigma)$

(ii) Soient $1 \leq k < n$ et $C \in \mathcal{C}_k$. On veut montrer que $T_n^{\mathrm{PDA}}(\varphi_{n,k}^{-1}(C))/T_n^{\mathrm{PDA}}(\mathcal{A}_{n,k}) = T_k(C)$. Il suffit de montrer que $\#\big(\varphi_{n,k}^{-1}(C)\big)/\#\mathcal{A}_{n,k} = 1/c_k$.
Se donner un cladogramme de $\mathcal{A}_{n,k}$ revient à :
- choisir $C_k \in \mathcal{C}_k$: c_k possibilités ;
- choisir $C_{n,k} \in \mathcal{C}_{n,k}$ (pour lequel on remplace les étiquettes $\{1,\dots,n-k\}$ des espèces par $\{k+1,\dots,n\}$) : c_{n-k} possibilités ;
- choisir l'arête de C_k à laquelle on fixe $\mathcal{C}_{n,k}$: $2k-1$ possibilités.

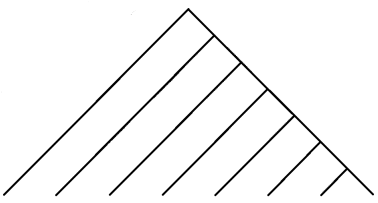
Donc $\#\mathcal{A}_{n,k} = c_k \cdot c_{n-k} \cdot (2k-1)$.
D'autre part, se donner un cladogramme de $\mathcal{A}_{n,k}$ revient à :
- choisir $C_{n,k} \in \mathcal{C}_{n,k}$ (pour lequel on remplace les étiquettes $\{1,\dots,n-k\}$ des espèces par $\{k+1,\dots,n\}$) : c_{n-k} possibilités ;
- choisir l'arête de C à laquelle on fixe $\mathcal{C}_{n,k}$: $2k-1$ possibilités.

Donc $\#\big(\varphi_{n,k}^{-1}(C)\big) = c_{n-k} \cdot (2k-1)$.
On en déduit alors que $\#\big(\varphi_{n,k}^{-1}(C)\big)/\#\mathcal{A}_{n,k} = 1/c_k$. ■

3.2 Le modèle du peigne

Définition 16

Le *modèle du peigne* a pour suite de distributions $(T_n^{\mathrm{p}})_{n\geq 1}$ où T_n^{p} est la distribution du cladogramme aléatoire à n feuilles qui a la forme suivante et dont les espèces sont réparties uniformément parmi les positions possibles.



Seules les deux dernières positions sont indifférenciées pour le cladogrammme, il y a donc, pour $n \geq 2$, $\binom{n}{2} \cdot (n-2)!$ positions différentes des espèces et on en déduit :

$$\forall n \geq 2, \forall C \in \mathcal{C}_n, T_n^{\mathrm{p}}(C) = \begin{cases} \frac{2}{n!} & \text{si } C \text{ a la forme d'un peigne} \\ 0 & \text{sinon} \end{cases}.$$

Proposition 17

La suite de distributions $(T_n^{\mathrm{p}})_{n\geq 1}$ sur les $\mathcal{C}_n$ du modèle du peigne vérifie les propriétés (i) et (ii).

Démonstration :

(i) $(T_n^{\mathrm{p}})_{n\geq 2}$ ne dépend que de la forme des cladogrammes donc vérifie (i).

(ii) Soient $1 \leq k < n$ et $C \in \mathcal{C}_k$. On veut montrer que $T_n^{\mathrm{p}}(\varphi_{n,k}^{-1}(C))/T_n^{\mathrm{p}}(\mathcal{A}_{n,k}) = T_k(C)$.

* Si C n'a pas la forme d'un peigne, alors soit $C' \in \varphi_{n,k}^{-1}(C)$, C' admet $\{k+1, \ldots, n\}$ comme clade et en le supprimant on obtient C donc C' n'a pas la forme d'un peigne (en supprimant un clade à un peigne, on conserve un peigne).
  Donc $T_n^{\mathrm{p}}(\varphi_{n,k}^{-1}(C))/T_n^{\mathrm{p}}(\mathcal{A}_{n,k}) = 0 = T_k(C)$.
* Supposons que C a la forme d'un peigne et que $k \leq n-2$.
  Les éléments de $\varphi_{n,k}^{-1}(C)$ qui ne sont pas de masse nulle sont les cladogrammes en forme de peigne, admettant $\{k+1, \ldots, n\}$ comme clade et dont C est le cladogramme obtenu en supprimant ce clade. Donc comme $n-k \geq 2$, ce sont les cladogrammes en forme de peigne, dont $\{k+1, \ldots, n\}$ sont les $n-k$ espèces à droite dans la représentation de la définition 14 et où les k espèces à droite sont celles de C (mis sous la forme de la définition 14) dans l'ordre, sauf pour les deux dernières, dont la position est indifférenciée pour C mais pas pour C'. Donc $\varphi_{n,k}^{-1}(C)$ contient exactement $\binom{n-k}{2}(n-k-2)! \cdot 2 = (n-k)!$ éléments en forme de peigne donc :

  $$T_n^{\mathrm{p}}(\varphi_{n,k}^{-1}(C)) = \frac{2 \cdot (n-k)!}{n!}.$$

  De même, les éléments de $\mathcal{A}_{n,k}$ qui ne sont pas de masse nulle sont les cladogrammes en forme de peigne, admettant $\{k+1, \ldots, n\}$ comme clade, c'est-à-dire dont les $n-k$ espèces à droite sont $\{k+1, \ldots, n\}$ dans la représentation de la définition 14 (donc dont les k espèces à gauche sont $\{1, \ldots, k\}$). Donc $\mathcal{A}_{n,k}$ contient exactement $\binom{n-k}{2}(n-k-2)! \cdot k! = (n-k)! \cdot k!/2$ éléments en forme de peigne donc :

  $$T_n^{\mathrm{p}}(\mathcal{A}_{n,k}) = \frac{(n-k)! \cdot k!}{n!}.$$

  Donc on a bien $T_n^{\mathrm{p}}(\varphi_{n,k}^{-1}(C))/T_n^{\mathrm{p}}(\mathcal{A}_{n,k}) = 2/k! = T_k(C)$.
* Supposons que C a la forme d'un peigne et que $k = n-1$.
  Alors tous les cladogrammes de $\mathcal{C}_n$ admettent $\{n\}$ comme sous-clade, donc $\mathcal{A}_{n,k} = \mathcal{C}_n$ et $T_n^{\mathrm{p}}(\mathcal{A}_{n,k}) = 1$.
  D'autre part, les éléments de $\varphi_{n,k}^{-1}(C)$ sont les cladogrammes obtenus en rajoutant l'espèce n à C (en rajoutant un nœud au milieu de l'une des $2k-1$

arêtes de C, dont celle rajoutée au-dessus de la racine). Le cladogramme obtenu a une forme de peigne si et seulement l'arête choisi et celle au-dessus de la racine ou l'une des arêtes de la diagonale supérieure droite de C (dans la représentation de la définition 14). Donc $\varphi_{n,k}^{-1}(C)$ contient exactement k éléments en forme de peigne donc, comme $k = n-1$, $T_n^{\mathrm{p}}(\varphi_{n,k}^{-1}(C)) = 2/(n-1)!$.
Donc on a bien $T_n^{\mathrm{p}}(\varphi_{n,k}^{-1}(C))/T_n^{\mathrm{p}}(\mathcal{A}_{n,k}) = 2/(n-1)! = T_k(C)$ (car $k = n-1$). ■

Le modèle du peigne correspond à un "déséquilibre maximal", au sens de la notion de déséquilibre définie en 4.1.

3.3 La conjecture d'Aldous

Comme aucun autre modèle dont la suite de distribution associée vérifierait (i) et (ii) n'a été construit, Aldous énonce dans [1] la conjecture suivante :

Conjecture 18

Les 3 suites de distributions du modèle ERM, du modèle PDA et du modèle du peigne sont les seules à vérifier (i) et (ii).

4 Le modèle Bêta

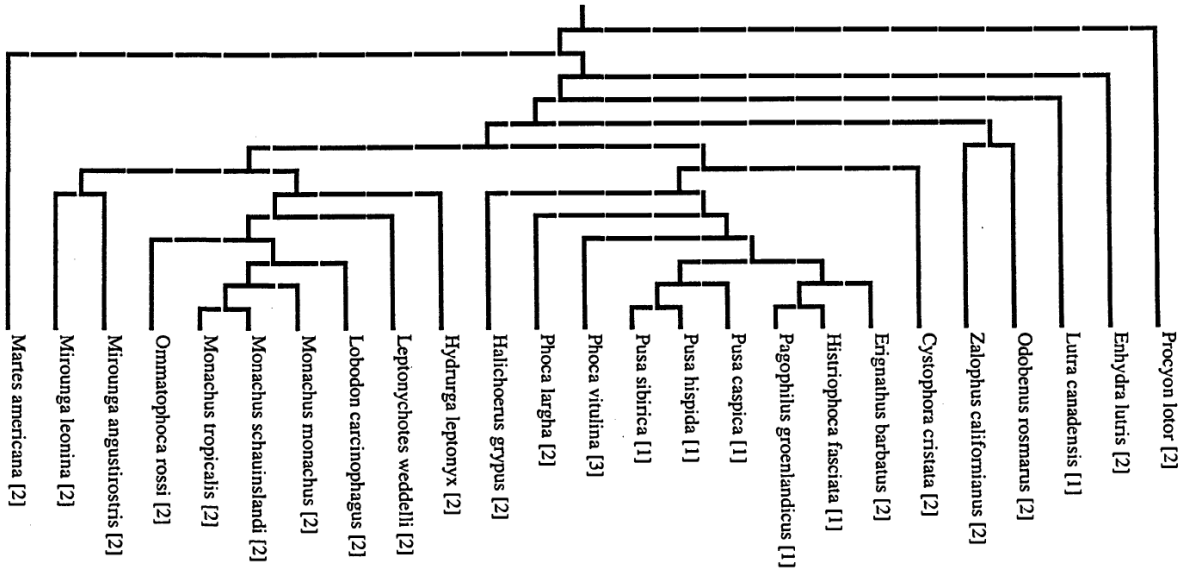
4.1 Pourquoi introduire ce modèle ?

D'après la conjecture 18, il semble donc que l'on ne peut pas trouver une famille de distributions à un paramètre caractérisant (i) et (ii). Cependant les trois modèles vérifiant (i) et (ii) ne sont pas satisfaisant car ils ne permettent pas de retrouver la forme des arbres phylogénétiques publiés dans TreeBASE.

Une manière de décrire cette forme est de s'intéresser au déséquilibre de l'arbre et plus précisément à la taille des deux sous-clades d'un nœud : à chaque nœud, on associe un couple (m, i), où m est la taille du clade associée à ce nœud et i est la taille du plus petit de ses deux sous-clades fils (donc $1 \leq i \leq m/2$).

Si i est proche de $m/2$, le nœud est équilibré et si i est proche de 1, il est déséquilibré. En particulier, la forme du peigne est donc bien la forme la plus déséquilibrée dans ce sens.

En observant les arbres phylogénétiques disponibles dans la base de donnée TreeBASE, on remarque que les nœuds sont en grande partie déséquilibrés, comme par exemple dans cet arbre ([2] Figure 2, issu de TreeBASE) représentant 25 espèces de phoques :



On obtient alors à partir de cet arbre le tableau suivant ([2] Table 2), regroupant les couples (m, i) pour tous les nœuds tels que $m \geq 6$ (si m est trop petit, "i proche de 1" et "i proche de $m/2$" n'ont pas de sens) :

Size of parent clade	**Size of smaller daughter clade**
25	1
24	1
23	1
22	1
21	2
19	9
10	1
9	2
9	1
8	1
7	1
7	1
6	3
6	1

On observe qu'à part les deux couples $(19, 9)$ et $(6, 3)$, tous les autres nœuds sont déséquilibrés.

On se demande alors si les modèles introduits précédemment fournissent des cladogrammes avec un déséquilibre similaire, comme le modèle de Yule pour les genres qui correspondait bien aux données. Cependant, comme on le constatera en 5, le modèle ERM fournit des cladogrammes plus équilibrés que ceux trouvés dans TreeBASE. En revanche, le modèle PDA fournit des arbres plus déséquilibrés. Ces deux modèles ne permettent donc pas de prédire la forme des arbres phylogénétiques.

On cherche donc une famille de distributions à un paramètre contenant ces deux modèles, qui permettrait donc peut-être d'obtenir un modèle "entre" les modèles ERM et PDA qui aurait un déséquilibre semblable à celui des données.

4.2 Modèles de branchements de Markov

Définition 19

Pour $n \geq 2$, soit q_n une distribution sur $[\![1, n-1]\!]$ symétrique (pour tout $1 \leq i \leq n-1$, $q_n(i) = q_n(n-i)$). Le *modèle de branchements de Markov* associé à la suite $(q_n)_{n\geq 2}$ est le modèle pour lequel un cladogramme aléatoire est obtenu par le procédé suivant :

- La racine d'un arbre aléatoire à n feuilles a i feuilles sur sa branche gauche et $n-i$ sur sa branche droite, où i est choisi aléatoirement avec pour loi q_n. Le choix des i espèces de la branche gauche est uniforme parmi les $\binom{n}{i}$ possibilités.
- On répète récursivement l'étape précédente sur chacune des deux branches.
- L'arbre obtenu s'interprète comme un cladogramme en ôtant les marqueurs gauche/droite.

Remarque : La suite de distribution $(T_n)_{n\geq 1}$ associée à un modèle de branchements de Markov vérifie (i) car, à chaque nœud, la répartition des espèces de chaque côté du nœud ne dépend pas des étiquettes des espèces.

Pour les modèles de branchements de Markov, comme dans le lemme 10 pour le modèle ERM, on obtient une relation entre la probabilité d'un cladogramme et les probabilités des deux sous-cladogrammes de sa racine, qui fournit un moyen assez simple de calculer la probabilité d'un cladogramme sous ce modèle :

Lemme 20

Soient $(q_n)_{n\geq 2}$ une suite de distributions symétriques sur les $[\![1, n-1]\!]$ et $(T_n)_{n\geq 1}$ la suite de distributions sur les $\mathcal{C}_n$ associée au modèle de branchements de Markov associé à $(q_n)_{n\geq 2}$.
Pour tous $n \geq 2$ et $C \in \mathcal{C}_n$, si k et $n-k$ sont les tailles des deux sous-clades de la racine de C et si $C_k \in \mathcal{C}_k$ et $C_{n-k} \in \mathcal{C}_{n-k}$ sont ces deux sous-clades, alors

$$T_n(C) = \frac{2 \cdot q_n(k)}{\binom{n}{k}} \cdot T_k(C_k) \cdot T_{n-k}(C_{n-k}).$$

Si pour N un nœud interne de C, on note (m_N, i_N) le couple correspondant à la taille du clade associé à N et la taille de son plus petit sous-clade, alors

$$T_n(C) = \prod_{N \text{ nœud de } C} \frac{2 \cdot q_{m_N}(i_N)}{\binom{m_N}{i_N}}.$$

Démonstration : Soient $n \geq 2$ et $C \in \mathcal{C}_n$, k et $n-k$ les tailles des deux sous-clades de la racine de C et $C_k \in \mathcal{C}_k$ et $C_{n-k} \in \mathcal{C}_{n-k}$ ces deux sous-clades. Il faut traiter les deux cas $k \neq n/2$ et $k = n/2$ séparément.

- Si $k \neq n/2$, la probabilité de C sous T_n est le produit de :
  - la probabilité que les deux sous-clades de la racine soient de tailles k et $n-k$: $q_n(k) + q_n(n-k)$ (car $k \neq n-k$ donc les deux choix sont bien distincts) ;

* la probabilité que les n espèces soient bien réparties dans les deux sous-clades : $1/\binom{n}{k}$;
* la probabilité que le sous-clade ayant k espèces soit C_k : $T_k(C_k)$;
* la probabilité que le sous-clade ayant $n-k$ espèces soit C_{n-k} : $T_{n-k}(C_{n-k})$.

Comme $q_n(k) = q_n(n-k)$, on a donc :

$$T_n(C) = \frac{2 \cdot q_n(k)}{\binom{n}{k}} \cdot T_k(C_k) \cdot T_{n-k}(C_{n-k}).$$

- Si $k = n/2$, les deux sous-clades de la racine sont de même taille donc le choix de quel sous-clade est C_k et lequel est C_{n-k} est arbitraire. Une fois ce choix fixé, la probabilité de C sous T_n est le produit de :
  * la probabilité que les deux sous-clades de la racine soient de tailles k et $n-k$: $q_n(n/2)$;
  * la probabilité que les n espèces soient bien réparties dans les deux sous-clades : $2/\binom{n}{n/2}$ (le facteur 2 provient du fait que les deux sous-clades sont encore indifférenciés puisqu'ils ont la même taille) ;
  * la probabilité que le sous-clade ayant les $n/2$ espèces de C_k soit C_k : $T_k(C_k)$;
  * la probabilité que le sous-clade ayant les $n/2$ espèces de C_{n-k} soit C_{n-k} : $T_{n-k}(C_{n-k})$.

  On a donc aussi :

$$T_n(C) = \frac{2 \cdot q_n(k)}{\binom{n}{k}} \cdot T_k(C_k) \cdot T_{n-k}(C_{n-k}).$$

On en déduit alors par une récurrence forte sur la taille des cladogrammes, l'expression de $T_n(C)$. ■

Définition 21

Soit $f :]0,1[\longrightarrow \mathbb{R}_+$ symétrique (pour tout $x \in]0,1[, f(x) = f(1-x)$) et telle que $\int_0^{1/2} xf(x)dx < +\infty$. Le *modèle de branchements de Markov associé à* f est le modèle de branchements de Markov associé à $(q_n)_{n\geq 2}$ avec :

$$\text{pour } 1 \leq i \leq n-1, q_n(i) = \frac{1}{a_n}\binom{n}{i}\int_0^1 x^i(1-x)^{n-i}f(x)dx$$

et où a_n est la constante normalisatrice :

$$a_n = \int_0^1 \left(1 - x^n - (1-x)^n\right) f(x)dx.$$

Démonstration : Vérifions que le modèle est bien défini.

- Par symétrie de f, comme $\int_0^{1/2} xf(x)dx < +\infty$, on a aussi $\int_{1/2}^1 (1-x)f(x)dx < +\infty$, pour $1 \leq i \leq n-1$, $x \longmapsto x^i(1-x)^{n-i}f(x)$ est intégrable sur $]0,1/2]$ (car $i \geq 1$) et sur $[1/2,1[$ (car $n-i \geq 1$). Donc q_n est bien définie.

- q_n est bien une mesure de probabilité :

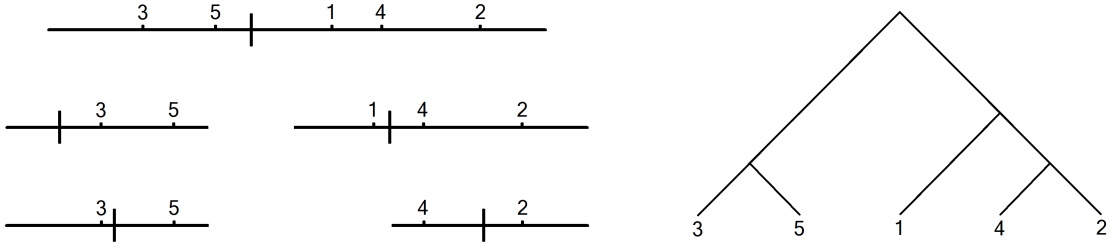
$$\begin{aligned}\sum_{i=1}^{n-1}\binom{n}{i}\int_0^1 x^i(1-x)^{n-i}f(x)dx &= \int_0^1\Big[\big(\sum_{i=0}^{n}\binom{n}{i}x^i(1-x)^{n-i}\big)-x^n-(1-x)^n\Big]f(x)dx\\ &= \int_0^1\big(1-x^n-(1-x)^n\big)f(x)dx\\ &= a_n.\end{aligned}$$

- q_n est symétrique car f l'est. ■

Proposition 22

Si f est la densité d'une probabilité sur $[0,1]$ (et que f est symétrique), alors le modèle de branchements de Markov associé à f s'interprète en termes de séparations d'intervalles :

- On commence avec n particules portant des étiquettes de 1 à n, chacune placée uniformément sur $[0,1]$.
- On sépare l'intervalle en un point aléatoire de $[0,1]$ choisi avec densité f.
- Pour chacun des deux intervalles, s'il contient au moins deux particules, on répète ce procédé, en séparant chaque intervalle $[a;b]$ en un point $a+X(b-a)$, où les X sont indépendants de densité f.
- Quand chaque intervalle ne contient plus qu'une particule, on interprète les séparations successive comme un cladogramme aléatoire, comme illustré dans la figure suivante.



On note que si, après la séparation d'un intervalle, l'un des deux sous-intervalles contient aucune particule, alors cette séparation est supprimée.

Pour démontrer cette proposition, nous allons utiliser le lemme suivant :

Lemme 23

Soient $a<b$ et $U_1,\dots,U_k$ des variables aléatoires indépendantes uniformes sur $[a,b]$. On pose $(T_1,\dots,T_k)$ le k-uplet ordonné par ordre croissant de $(U_1,\dots,U_k)$: il existe ς une variable aléatoire à valeurs dans $\mathcal{S}_k$ telle que pour $1\le i\le k$, $T_i=U_{\varsigma(i)}$ et pour $i\le j$, $T_i\le T_j$. Alors la loi de $(T_1,\dots,T_k)$ est de densité p où :

$$\forall(t_1,\dots,t_k)\in\mathbb{R}^k, p(t_1,\dots,t_k)=\frac{k!}{(b-a)^k}\cdot\mathbb{1}_{a\le t_1<\dots<t_k\le b}.$$

En particulier, on a l'égalité :

$$\int_{[a,b]^k}\mathbb{1}_{t_1<\dots<t_k}dt_1\cdots dt_k=\frac{(b-a)^k}{k!}.$$

Démonstration du lemme : Pour $\sigma \in \mathcal{S}_k$, on pose $A_\sigma = \{U_{\sigma(1)} < \cdots < U_{\sigma(k)}\}$ et $B = \{\exists i \neq j : U_i = U_j\}$. Alors on a :

$$\Omega = B \sqcup \bigsqcup_{\sigma \in \mathcal{S}_k} A_\sigma \quad \text{et} \quad \mathbb{P}(B) = 0.$$

En effet, pour $i \neq j$, $\mathbb{P}(U_i = U_j) = P_{(U_i,U_j)}\big(\{(x,x) \mid x \in [a,b]\}\big) = 0$ car U_i et U_j sont indépendantes et à densité et que $\lambda\big(\{(x,x) \mid x \in [a,b]\}\big) = 0$.

Soit $\varphi : \mathbb{R}^k \longrightarrow \mathbb{R}$ continue et bornée,

$$\begin{aligned}
\mathbb{E}[\varphi(T_1,\ldots,T_k)] &= \sum_{\sigma\in\mathcal{S}_k} \mathbb{E}[\varphi(T_1,\ldots,T_k)\mathbb{1}_{A_\sigma}] \\
&= \sum_{\sigma\in\mathcal{S}_k} \mathbb{E}[\varphi(U_{\sigma(1)},\ldots,U_{\sigma(k)})\mathbb{1}_{U_{\sigma(1)}<\cdots<U_{\sigma(k)}}] \\
&= \sum_{\sigma\in\mathcal{S}_k} \int_{[a,b]^k} \varphi(u_{\sigma(1)},\ldots,u_{\sigma(k)})\mathbb{1}_{u_{\sigma(1)}<\cdots<u_{\sigma(k)}} \cdot \frac{1}{(b-a)^k} \cdot du_1\cdots du_k \\
&= \sum_{\sigma\in\mathcal{S}_k} \int_{[a,b]^k} \varphi(v_1,\ldots,v_k)\mathbb{1}_{v_1<\cdots<v_k} \cdot \frac{1}{(b-a)^k} \cdot dv_1\cdots dv_k \\
&= k! \int_{\mathbb{R}^k} \varphi(v_1,\ldots,v_k)\mathbb{1}_{a\leq v_1<\cdots<v_k\leq b} \cdot \frac{1}{(b-a)^k} \cdot dv_1\cdots dv_k \\
&= \int_{\mathbb{R}^k} \varphi(v_1,\ldots,v_k)\cdot p(v_1,\ldots,v_k)\cdot dv_1\cdots dv_k.
\end{aligned}$$

Donc $(T_1,\ldots,T_k)$ est de densité p. L'égalité suivante s'obtient à partir de $\int_{\mathbb{R}^k} p = 1$. ■

Démonstration de la proposition : Les n particules placées dans $[0,1]$ correspondent à $U_1,\ldots,U_n$ des variables aléatoires indépendantes uniformes sur $[0,1]$. On pose $(T_1,\ldots,T_n)$ le n-uplet ordonné par ordre croissant de $(U_1,\ldots,U_n)$.

On note X la variable aléatoire correspondant à la position du point de séparation. Elle suit la loi de densité f.

X et $(T_1,\ldots,T_n)$ sont indépendants donc la loi de $(X,T_1,\ldots,T_n)$ a pour densité $(t_1,\ldots,t_k) \longmapsto f(x)\mathbb{1}_{0\leq x\leq 1}\cdot n!\ \mathbb{1}_{0\leq t_1<\cdots<t_n\leq 1}$

- Soit $1 \leq i \leq n-1$. La probabilité de séparer les n espèces en un groupe de i espèces à gauche et un groupe de $n-i$ espèces à droite est la probabilité que $T_i \leq X < T_{i+1}$ sachant que $T_1 \leq X < T_n$ (sinon le tirage de la séparation est supprimé). En utilisant le lemme 23 :

$$\begin{aligned}
\mathbb{P}(T_i \leq X < T_{i+1}) &= \int_{\mathbb{R}^{n+1}} \mathbb{1}_{t_i\leq x<t_{i+1}} \cdot f(x)\mathbb{1}_{0\leq x\leq 1}\cdot n!\ \mathbb{1}_{0\leq t_1<\cdots<t_n\leq 1} dx dt_1\cdots dt_n \\
&= \int_0^1 n!\Big(\int_{[0,x]^i} \mathbb{1}_{t_1<\cdots<t_i} dt_1\cdots dt_i\Big)\Big(\int_{[x,1]^{n-i}} \mathbb{1}_{t_{i+1}<\cdots<t_n} dt_{i+1}\cdots dt_n\Big) f(x)dx \\
&= \int_0^1 n!\cdot\frac{x^i}{i!}\cdot\frac{(1-x)^{n-i}}{(n-i)!}\cdot f(x)dx \\
&= \binom{n}{i}\int_0^1 x^i(1-x)^{n-i} f(x)dx \\
&= a_n q_n(i)
\end{aligned}$$

$$\text{et } \mathbb{P}(T_1 \leq X < T_n) = \sum_{i=1}^{n-1} \mathbb{P}(T_i \leq X < T_{i+1}) = a_n \sum_{i=1}^{n-1} q_n(i) = a_n.$$

Donc la probabilité de séparer les n espèces en un groupe de i espèces à gauche et un groupe de $n-i$ espèces à droite est :

$$\mathbb{P}(T_i \leq X < T_{i+1} \mid T_1 \leq X < T_n) = \frac{\mathbb{P}(T_i \leq X < T_{i+1})}{\mathbb{P}(T_1 \leq X < T_n)} = q_n(i).$$

- Soient $1 \leq i \leq n$ et I une partie de $[\![1, n]\!]$ à i éléments. Sachant que $T_i \leq X < T_{i+1}$, on cherche la probabilité que les i espèces à gauche de la séparation portent les étiquettes de I, c'est-à-dire la probabilité que $\varsigma(I) = [\![1, i]\!]$, avec ς défini comme dans le lemme 23.
  On pose $\mathcal{G} = \{\sigma \in \mathcal{S}_n \mid \sigma(I) = [\![1, i]\!]\}$. On a les inclusions :

$$B \sqcup \bigsqcup_{\sigma \in \mathcal{G}} A_\sigma \subset \{\varsigma(I) = [\![1, i]\!]\} \subset \bigsqcup_{\sigma \in \mathcal{G}} A_\sigma.$$

  Comme $\mathbb{P}(B) = 0$, on en déduit :

$$\begin{aligned}\mathbb{P}(\varsigma(I) = [\![1, i]\!] \mid T_i \leq X < T_{i+1}) &= \sum_{\sigma \in \mathcal{G}} \mathbb{E}[\mathbb{1}_{A_\sigma} \mid T_i \leq X < T_{i+1}] \\ &= \sum_{\sigma \in \mathcal{G}} \frac{\mathbb{E}[\mathbb{1}_{U_{\sigma(1)} < \cdots < U_{\sigma(i)} \leq X < U_{\sigma(i+1)} < \cdots < U_{\sigma(n)}}]}{\mathbb{P}(T_i \leq X < T_{i+1})}\end{aligned}$$

  Or pour $\sigma \in \mathcal{S}_n$, en effectuant le changement de variables $(v_1, \ldots, v_n) = (u_{\sigma(1)}, \ldots, u_{\sigma(n)})$,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{U_{\sigma(1)} < \cdots < U_{\sigma(i)} \leq X < U_{\sigma(i+1)} < \cdots < U_{\sigma(n)}}] &= \int_{\mathbb{R}^{n+1}} \mathbb{1}_{0 \leq v_1 < \cdots < v_i \leq x < v_{i+1} < \cdots < v_n \leq 1} f(x) \cdot dx dv_1 \cdots dv_n \\ &= \frac{1}{n!} \cdot \mathbb{P}(T_i \leq X < T_{i+1})\end{aligned}$$

  d'après le calcul de $\mathbb{P}(T_i \leq X < T_{i+1}$ effectué au point précédent.
  On a donc $\mathbb{P}(\varsigma(I) = [\![1, i]\!] \mid T_i \leq X < T_{i+1}) = \#\mathcal{G}/n!$. Or se donner un élément de $\mathcal{G}$ revient à se donner une bijection de I dans $[\![1, i]\!]$ et une bijection de $I \smallsetminus [\![1, i]\!]$ dans $[\![i+1, n]\!]$, donc $\#\mathcal{G} = i! \cdot (n-i)!$. On obtient ainsi $\mathbb{P}(\varsigma(I) = [\![1, i]\!] \mid T_i \leq X < T_{i+1}) = 1/\binom{k}{i}$. Le choix de l'ensemble I des étiquettes des i espèces dans le groupe de gauche est donc uniforme sur l'ensemble des parties à i éléments de $[\![1, n]\!]$.
- On a montré que la première étape des deux procédés est identique : séparer les n espèces en 2 groupes avec la probabilité $q_n(i)$ que le groupe de gauche ait i espèces et ensuite avec un choix uniforme de ces i espèces. Le procédé de branchement de Markov consistant à répéter récursivement cette étape à chaque groupe, il suffit de vérifier que le procédé de séparations d'intervalles fait de même.
  Pour cela il suffit de montrer que si un intervalle obtenu après séparation contient i espèces, alors ces i particules suivent la loi d'un i-uplet de variables aléatoires indépendantes uniformes à valeurs dans cet intervalle.
  Fixons $1 \leq i \leq n-1$. On se place sous la condition $\{T_i \leq X < T_{i+1}\}$ et on note alors $\tilde{\mathbb{P}}$ et $\tilde{\mathbb{E}}$ la probabilité et la fonction espérance associées à ce conditionnement. Montrons que sous $\tilde{\mathbb{P}}$, la loi conditionnelle de $(T_1, \ldots, T_i)$ sachant X est la loi d'un i-uplet ordonné de variables aléatoires indépendantes uniformes à valeur dans $[0, X]$.

Déterminons d'abord la loi de $(X, T_1, \dots, T_n)$ sous $\tilde{\mathbb{P}}$. Soit $\varphi : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}_+$,

$$\begin{aligned}\mathbb{E}[\varphi(X, T_1, \dots, T_n) \mid T_i \leq X < T_{i+1}] &= \frac{\mathbb{E}[\varphi(X, T_1, \dots, T_n)\mathbb{1}_{T_i \leq X < T_{i+1}}]}{\mathbb{P}(T_i \leq X < T_{i+1})} \\ &= \frac{1}{a_n q_n(i)} \cdot \int_{\mathbb{R}^{n+1}} \varphi(x, t_1, \dots, t_n) f(x) \\ &\qquad \cdot n!\ \mathbb{1}_{0 \leq t_1 < \dots t_i \leq x < t_{i+1} < \dots < t_n \leq 1} dx dt_1 \cdots dt_n.\end{aligned}$$

Donc la loi de $(X, T_1, \dots, T_n)$ sous $\tilde{\mathbb{P}}$ est la loi de densité p_1 où :

$$p_1(x, t_1, \dots, t_n) = \frac{n!}{a_n q_n(i)} \cdot f(x) \cdot \mathbb{1}_{0 \leq t_1 < \dots t_i \leq x < t_{i+1} < \dots < t_n \leq 1}.$$

On en déduit alors que la loi de $(X, T_1, \dots, T_i)$ sous $\tilde{\mathbb{P}}$ est la loi de densité p_2 où :

$$\begin{aligned}p_2(x, t_1, \dots, t_i) &= \int_{\mathbb{R}^{n-i}} p_1(x, t_1, \dots, t_n) dt_{i+1} \cdots dt_n \\ &= \frac{n!}{a_n q_n(i)} \cdot f(x) \cdot \mathbb{1}_{0 \leq t_1 < \dots < t_i \leq x \leq 1} \int_{[x,1]^{n-i}} \mathbb{1}_{t_{i+1} < \dots < t_n} dt_{i+1} \cdots dt_n \\ &= \frac{n!}{a_n q_n(i)} \cdot f(x) \cdot \mathbb{1}_{0 \leq t_1 < \dots < t_i \leq x \leq 1} \cdot \frac{(1-x)^{n-i}}{(n-i)!}.\end{aligned}$$

Et que la loi de X sous $\tilde{\mathbb{P}}$ est la loi de densité p_3 où :

$$p_3(x) = \int_{\mathbb{R}^i} p_2(x, t_1, \dots, t_i) dt_1 \cdots dt_i = \frac{n!}{a_n q_n(i)} \cdot f(x) \cdot \frac{(1-x)^{n-i}}{(n-i)!} \cdot \frac{x^i}{i!} \mathbb{1}_{0 \leq x \leq 1}.$$

Déterminons maintenant, sous $\tilde{\mathbb{P}}$, la loi conditionnelle de $(T_1, \dots, T_i)$ sachant X. Soient $\varphi : \mathbb{R}^i \longrightarrow \mathbb{R}_+$ et $\psi : \mathbb{R} \longrightarrow \mathbb{R}_+$,

$$\begin{aligned}\tilde{\mathbb{E}}[\varphi(T_1, \dots, T_n)\psi(X)] &= \int_{\mathbb{R}^{i+1}} \varphi(t_1, \dots, t_i)\psi(x) \cdot p_2(x, t_1, \dots, t_i) dx dt_1 \cdots dt_i \\ &= \int_0^1 \left(\int_{\mathbb{R}^i} \varphi(t_1, \dots, t_i) \mathbb{1}_{0 \leq t_1 < \dots < t_i \leq x} dt_1 \cdots dt_i \right) \\ &\qquad \cdot \psi(x) \cdot \frac{n!}{a_n q_n(i)} \cdot f(x) \cdot \frac{(1-x)^{n-i}}{(n-i)!} dx \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}^i} \varphi(t_1, \dots, t_i) \mathbb{1}_{0 \leq t_1 < \dots < t_i \leq x} \frac{i!}{x^i} dt_1 \cdots dt_i \right) \cdot \psi(x) p_3(x) dx \\ &= \tilde{\mathbb{E}}\left[\psi(X) \int_{\mathbb{R}^i} \varphi(t_1, \dots, t_i) \mathbb{1}_{0 \leq t_1 < \dots < t_i \leq X} \frac{i!}{X^i} dt_1 \cdots dt_i \right].\end{aligned}$$

Donc, pour tout $\varphi : \mathbb{R}^i \longrightarrow \mathbb{R}_+$, comme la variable aléatoire de droite dans l'inégalité ci-dessous est bien $\sigma(X)$-mesurable,

$$\tilde{\mathbb{E}}\left[\varphi(T_1, \dots, T_i) \mid X\right] = \int_{\mathbb{R}^i} \varphi(t_1, \dots, t_i) \mathbb{1}_{0 \leq t_1 < \dots < t_i \leq X \leq 1} \frac{i!}{X^i} dt_1 \cdots dt_i.$$

Donc sous $\tilde{\mathbb{P}}$, la loi conditionnelle de $(T_1, \dots, T_i)$ sachant X est la loi de densité :

$$(t_1, \dots, t_i) \longmapsto \mathbb{1}_{0 \leq t_1 < \dots < t_i \leq X \leq 1} \frac{i!}{X^i}.$$

C'est donc bien la loi d'un i-uplet ordonné de variables aléatoires indépendantes uniformes à valeur dans $[0, X]$.
Et, par symétrie, sous $\tilde{\mathbb{P}}$, la loi conditionnelle de $(T_{i+1}, \ldots, T_n)$ sachant X est la loi d'un i-uplet ordonné de variables aléatoires indépendantes uniformes à valeur dans $[X, 1]$. ■

Remarque : Dans le cas où f est la densité d'une probabilité sur $[0, 1]$, on a, en utilisant la symétrie de f :

$$\begin{aligned} a_n &= \int_0^1 \big(1 - x^n - (1-x)^n\big) f(x) dx \\ &= \int_0^1 f(x) dx - \int_0^1 x^n f(x) dx - \int_0^1 x^n f(1-x) dx \\ &= 1 - 2\int_0^1 x^n f(x) dx. \end{aligned}$$

4.3 Définition du modèle Bêta

Définition 24

Le *modèle Bêta* est le modèle associé à la famille $(T_n^\beta)_{n\geq 2}$ paramétrisée par $2 \leq \beta \leq +\infty$ et définie ainsi :

- Pour $-1 < \beta < +\infty$, $(T_n^\beta)_{n\geq 2}$ est la suite de distribution obtenue par le procédé de séparations d'intervalles où f est la densité de la loi Beta$(\beta+1, \beta+1)$, c'est-à-dire :
$$\text{pour } 0 < x < 1, f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \cdot x^\beta (1-x)^\beta.$$
- Pour $\beta = +\infty$, $(T_n^\beta)_{n\geq 2}$ est obtenue par le procédé de séparations d'intervalles où les intervalles sont séparés de manière déterministe en leur milieu.
- Pour $-2 < \beta \leq -1$, $(T_n^\beta)_{n\geq 2}$ est obtenue par le modèle de branchements de Markov associé à f où :
$$\text{pour } 0 < x < 1, f(x) = x^\beta (1-x)^\beta.$$
- Pour $\beta = -2$, $(T_n^\beta)_{n\geq 2}$ est la suite $(T_n^{\mathrm{p}})_{n\geq 2}$ obtenue par le modèle du peigne.

Démonstration :

- Pour $-1 < \beta < +\infty$, f est symétrique et c'est bien la densité d'une probabilité :
$$\int_0^1 f(x) dx = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \cdot \int_0^1 x^\beta (1-x)^\beta dx = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \cdot \mathrm{B}(\beta+1, \beta+1) = 1.$$
- Pour $-2 < \beta < -1$, f est symétrique, et comme $\beta > -2$, on a bien :
$$\int_0^{1/2} x f(x) dx = \int_0^{1/2} x^{\beta+1} (1-x)^\beta dx < +\infty.$$

Donc le modèle est bien défini. ■

Ce modèle est introduit par Aldous dans [1], où il l'appelle *the beta-splitting model*.

Remarque :

- Pour $-1 < \beta < +\infty$, $(T_n^\beta)_{n\geq 2}$ est également obtenue par le modèle de branchements de Markov associé à $(q_n^\beta)_{n\geq 2}$ avec, pour $1 \leq i \leq n-1$:

$$\begin{aligned} q_n^\beta(i) &= \frac{1}{a_n^\beta} \cdot \binom{n}{i} \int_0^1 \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} x^{\beta+i}(1-x)^{\beta+n-i} dx \\ &= \frac{1}{a_n^\beta} \cdot \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)} \cdot \frac{\Gamma(n+1)}{\Gamma(i+1)\Gamma(n-i+1)} \cdot \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(2\beta+n+2)} \\ &= \frac{1}{\alpha_n(\beta)} \cdot \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)} \end{aligned}$$

$$\text{où } \alpha_n(\beta) = \left(1 - 2 \cdot \frac{\Gamma(\beta+n+1)\Gamma(\beta+1)}{\Gamma(2\beta+n+2)}\right) \cdot \frac{\Gamma^2(\beta+1)\Gamma(2\beta+n+2)}{\Gamma(n+1)\Gamma(2\beta+2)}$$

  en utilisant $a_n^\beta = 1 - 2\int_0^1 x^n f(x)dx = 1 - 2\mathrm{B}(\beta+n+1, \beta+1)$.
- Le modèle Bêta pour $\beta = +\infty$ se comprend comme la limite des modèles Bêta pour $\beta \longrightarrow +\infty$. En effet, quand $\beta \longrightarrow +\infty$, on a :

$$\begin{aligned} q_n^\beta(i) &= \frac{1}{a_n^\beta}\binom{n}{i} \cdot \frac{\Gamma(2\beta+2)}{\Gamma(2\beta+n+2)} \cdot \frac{\Gamma(\beta+i+1)}{\Gamma(\beta+1)} \cdot \frac{\Gamma(\beta+n-i+1)}{\Gamma(\beta+1)} \\ &\sim \frac{1}{a_n^\beta}\binom{n}{i} \cdot \frac{1}{(2\beta)^n} \cdot \beta^i \cdot \beta^{n-i} \\ \text{et } a_n^\beta &= 1 - 2 \cdot \frac{\Gamma(\beta+n+1)\Gamma(\beta+1)}{\Gamma(2\beta+n+2)} \longrightarrow 1 - 2 \cdot \frac{1}{2^n} \\ \text{donc } q_n^\beta(i) &\longrightarrow \frac{1}{2^n-2}\binom{n}{i}. \end{aligned}$$

  Posons $q_n(i) = \binom{n}{i}/(2^n-2)$ et montrons que le modèle AB pour $\beta = +\infty$ est bien le modèle de branchements de Markov associé à $(q_n)_{n\geq 2}$.
  Il suffit de montrer que, si $U_1, \ldots, U_n$ sont des variables aléatoires indépendantes uniformes sur $[0,1]$, alors $q_n(i)$ est bien la probabilité qu'exactement i de ces variables soient dans $[0, 1/2]$ sachant qu'il n'y en a ni 0 ni n (car sinon la séparation de l'intervalle est annulée). En effet, avec les notations du lemme 23 :

$$\mathbb{P}(T_i \leq \frac{1}{2} < T_{i+1} \mid T_1 \leq \frac{1}{2} \text{ et } T_n > \frac{1}{2}) = \frac{\mathbb{P}(T_i \leq \frac{1}{2} < T_{i+1})}{\mathbb{P}(T_1 \leq \frac{1}{2} \text{ et } T_n > \frac{1}{2})}$$

$$\begin{aligned}
\text{or } \mathbb{P}(T_i \leq \frac{1}{2} < T_{i+1}) &= \int_{[0,1]^n} \mathbb{1}_{t_i \leq \frac{1}{2} < t_{i+1}} \cdot n! \cdot \mathbb{1}_{t_1 < \cdots < t_n} dt_1 \cdots dt_n \\
&= n! \left(\int_{[0,1/2]^i} \mathbb{1}_{t_1 < \cdots < t_i} dt_1 \cdots dt_i \right) \left(\int_{[1/2,1]^{n-i}} \mathbb{1}_{t_{i+1} < \cdots < t_n} dt_1 \cdots dt_n \right) \\
&= n! \cdot \frac{(1/2)^i}{i!} \cdot \frac{(1/2)^{n-i}}{(n-i)!} \\
&= \binom{n}{i} \cdot \frac{1}{2^n}
\end{aligned}$$

$$\begin{aligned}
\text{et } \mathbb{P}(T_1 \leq \frac{1}{2} \text{ et } T_n > \frac{1}{2}) &= \int_{[0,1]^n} \mathbb{1}_{t_1 \leq \frac{1}{2}} \mathbb{1}_{t_n > \frac{1}{2}} \cdot n! \cdot \mathbb{1}_{t_1 < \cdots < t_n} dt_1 \cdots dt_n \\
&= n! \int_0^{1/2} \int_{1/2}^1 \left(\int_{[t_1, t_n]^{n-2}} \mathbb{1}_{t_2 < \cdots < t_{n-2}} dt_2 \cdots dt_{n-2} \right) dt_n dt_1 \\
&= n! \int_0^{1/2} \int_{1/2}^1 \frac{(t_n - t_1)^{n-2}}{(n-2)!} dt_n dt_1 \\
&= n \int_0^{1/2} (1 - t_1)^{n-1} - (\frac{1}{2} - t_1)^{n-1} dt_1 \\
&= -\frac{1}{2^n} + 1 - \frac{1}{2^n} = 1 - \frac{1}{2^{n-1}}.
\end{aligned}$$

Donc on a bien :

$$\mathbb{P}(T_i \leq \frac{1}{2} < T_{i+1} \mid T_1 \leq \frac{1}{2} \text{ et } T_n > \frac{1}{2}) = q_n(i).$$

- Pour $-2 < \beta \leq -1$, la fonction f n'est pas intégrable sur $[0,1]$ et donc elle ne peut pas être normalisée pour obtenir une densité de mesure de probabilité. Par définition, $(T_n^\beta)_{n \geq 2}$ est obtenue par le modèle de branchements de Markov associé à $(q_n^\beta)_{n \geq 2}$ avec

$$\begin{aligned}
q_n^\beta(i) &= \frac{1}{a_n^\beta} \cdot \binom{n}{i} \int_0^1 x^{\beta+i} (1-x)^{\beta+n-i} dx \\
&= \frac{1}{a_n^\beta} \cdot \frac{\Gamma(n+1)}{\Gamma(i+1)\Gamma(n-i+1)} \cdot \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(2\beta+n+2)} \\
&= \frac{1}{\alpha_n(\beta)} \cdot \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)} \\
&\qquad \text{où } \alpha_n(\beta) = a_n^\beta \cdot \frac{\Gamma(2\beta+n+2)}{\Gamma(n+1)}.
\end{aligned}$$

Mais dans ce cas on ne peut pas obtenir a_n^β de manière "explicite" car f n'est pas une densité de probabilité. Comme $\alpha_n(\beta)$ est une constante normalisatrice, on a cependant cette relation (vraie aussi dans le cas $-1 < \beta < +\infty$) :

$$\alpha_n(\beta) = \sum_{i=1}^{n-1} \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)}.$$

- Le modèle Bêta pour $\beta = -2$ (qui est le modèle du peigne) se comprend comme la limite des modèles Bêta pour $\beta \longrightarrow -2^+$.
  Pour $2 \leq i \leq n-2$, quand $\beta \longrightarrow -2^+$,

$$\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)} \longrightarrow \frac{\Gamma(i-1)\Gamma(n-i-1)}{\Gamma(i+1)\Gamma(n-i+1)}.$$

  Mais, comme $\Gamma(t) \sim 1/t$ quand $t \longrightarrow 0^+$, on a pour $i = 1$ ou $i = n-1$, quand $\beta \longrightarrow -2^+$,

$$\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)} \sim \frac{1}{\beta+2} \cdot \frac{\Gamma(n-2)}{\Gamma(2)\Gamma(n)} \longrightarrow +\infty.$$

  Donc on en déduit :

$$\alpha_n(\beta) \sim \frac{2}{\beta+2} \cdot \frac{\Gamma(n-2)}{\Gamma(2)\Gamma(n)} \text{ et } q_n^{\beta}(i) \longrightarrow q_n^{\mathrm{p}}(i) := \begin{cases} \frac{1}{2} \text{ si } i = 1 \text{ ou } i = n-1 \\ 0 \text{ sinon} \end{cases}.$$

  Et le modèle de branchements de Markov associé à la suite $(q_n^{\mathrm{p}})_{n\leq 2}$ est bien le modèle du peigne puisque pour un clade de n espèces, les deux sous-clades sont forcément de tailles 1 et $n-1$.

4.4 Cas particuliers

La famille à un paramètre formée par le modèle Bêta pour $-2 \leq \beta \leq +\infty$ contient le modèle du peigne mais aussi les modèle ERM et PDA, comme le montre la proposition suivante.

Proposition 25

- Pour $\beta = 0$, on retrouve le modèle ERM $(T_n^0)_{n\geq 1} = (T_n^{\mathrm{ERM}})_{n\geq 1}$. On a pour $1 \leq i \leq n-1$:

$$q_n^{\mathrm{ERM}}(i) = \frac{2}{n-1}.$$

- Pour $\beta = -3/2$, on retrouve le modèle PDA : $(T_n^{-3/2})_{n\geq 1} = (T_n^{\mathrm{PDA}})_{n\geq 1}$. On a pour $1 \leq i \leq n-1$:

$$q_n^{\mathrm{PDA}}(i) = \binom{n}{i} \frac{c_i c_{n-i}}{c_n}.$$

- Pour $\beta = -1$, le modèle Bêta est appelé *modèle AB* (pour *Aldous Branching*). C'est le modèle de branchements de Markov associé à la suite $(q_n^{\mathrm{AB}})_{n\geq 2}$ vérifiant pour $1 \leq i \leq n-1$:

$$q_n^{\mathrm{AB}}(i) = \frac{n}{2h_{n-1}} \cdot \frac{1}{i(n-i)}$$

  où h_n est la somme harmonique : $h_n = \sum_{i=1}^{n} \frac{1}{n}$.

Le modèle AB n'a pas été rencontré précédemment, mais nous verrons dans la partie 5 que ce cas particulier est intéressant car il fournit des cladogrammes assez proches de ceux trouvés dans TreeBASE.

Démonstration :

- Pour $\beta = 0$, pour $1 \leq i \leq n-1$:

$$q_n^0(i) = \frac{1}{\alpha_n(0)} \cdot \frac{\Gamma(0+i+1)\Gamma(0+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)} = \frac{1}{\alpha_n(0)} = \frac{1}{n-1}$$

$$\begin{aligned}\text{car } \alpha_n(0) &= \left(1 - 2 \cdot \frac{\Gamma(0+n+1)\Gamma(0+1)}{\Gamma(2\cdot 0+n+2)}\right) \cdot \frac{\Gamma^2(0+1)\Gamma(2\cdot 0+n+2)}{\Gamma(n+1)\Gamma(2\cdot 0+2)} \\ &= \left(1 - \frac{2}{n+1}\right) \cdot (n+1) \\ &= n-1.\end{aligned}$$

Soit $C \in \mathcal{C}_n$, d'après les lemmes 20 et 10,

$$T_n^0(C) = \prod_{N \text{ nœud de } C} \frac{2 \cdot q_{m_N}^0(i_N)}{\binom{m_N}{i_N}} = \prod_{N \text{ nœud de } C} \frac{2}{(m_N-1)\binom{m_N}{i_N}} = T_n^{\mathrm{ERM}}(C).$$

Donc on retrouve bien le modèle ERM.
Les q_n^0 sont les distributions uniformes sur les $[\![1, n-1]\!]$, ce qui explique le nom ERM (Equal Rate Markov), car c'est le procédé de branchements de Markov pour lequel toutes les tailles possibles pour le sous-clade gauche d'un nœud sont équiprobables.

- Pour $\beta = -3/2$, montrons que $(T_n^{-3/2})_{n\geq 1} = (T_n^{\mathrm{PDA}})_{n\geq 1}$.
On rappelle que c_n est le nombre de cladogrammes à n feuilles.
Soit $n \geq 2$ et $1 \leq i < n/2$. Se donner un cladogramme C à n feuilles dont la racine a des clades fils de tailles i et $n-i$ revient à :
  - choisir i espèces parmi les n espèces pour les mettre dans le sous-clade de taille i : $\binom{n}{i}$ possibilités ;
  - choisir un cladogramme sur ces i espèces (le cladogramme du sous-clade de taille i) : c_i possibilités ;
  - choisir un cladogramme sur les $n-i$ autres espèces (le cladogramme de l'autre sous-clade) : c_{n-i} possibilités.

Donc, si on note $\mathcal{B}_{n,i}$ l'ensemble des cladogrammes à n feuilles dont la racine a des sous-clades fils de tailles i et $n-i$, alors pour $i < n/2$:

$$\#\mathcal{B}_{n,i} = \binom{n}{i} c_i c_{n-i}.$$

Si n est pair et que $i = n/2$, alors en procédant de même pour dénombrer $\mathcal{B}_{n,n/2}$, on compte exactement 2 fois chaque cladogramme, donc :

$$\#\mathcal{B}_{n,i} = \frac{1}{2}\binom{n}{i} c_i c_{n-i}.$$

Comme les $\mathcal{B}_{n,i}$ pour $1 \leq i \leq \lfloor n/2 \rfloor$ forment une partition de $\mathcal{C}_n$, on en déduit :

$$c_n = \sum_{i=1}^{\lfloor n/2 \rfloor} \#\mathcal{B}_{n,i} = \sum_{i=1}^{n} \frac{1}{2}\binom{n}{i} c_i c_{n-i}$$

car dans la somme de droite, tous les termes apparaissent deux fois sauf celui pour $i = n/2$ (quand n est pair).
D'autre part, pour $n \geq 2$ on a :

$$c_n = (2n-3)!! = \frac{2^{n-1}}{\Gamma(1/2)} \cdot \Gamma(1/2) \cdot \frac{1}{2} \cdot \frac{3}{2} \cdots \left(n - \frac{3}{2}\right) = \frac{2^{n-1}}{\Gamma(1/2)} \cdot \Gamma\left(n - \frac{1}{2}\right).$$

Et donc pour $1 \leq i \leq n-1$:

$$\begin{aligned} q_n^{-3/2}(i) &= \frac{1}{\alpha_n(-3/2)} \cdot \frac{\Gamma(-\frac{3}{2}+i+1)\Gamma(-\frac{3}{2}+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)} \\ &= \frac{1}{\alpha_n(-3/2)} \cdot \binom{n}{i} \cdot \frac{1}{n!} \cdot \Gamma(i-\frac{1}{2}) \cdot \Gamma(n-i-\frac{1}{2}) \\ &= \frac{1}{\alpha_n(-3/2)} \cdot \binom{n}{i} \cdot \frac{1}{n!} \cdot \frac{c_i \Gamma(1/2)}{2^{i-1}} \cdot \frac{c_{n-i}\Gamma(1/2)}{2^{n-i-1}} \\ &= \frac{1}{\alpha_n(-3/2)} \cdot \binom{n}{i} \cdot \frac{\Gamma(1/2)^2}{2^n n!} \cdot c_i c_{n-i} \\ &= \binom{n}{i} \cdot \frac{c_i c_{n-i}}{2c_n} \end{aligned}$$

$$\text{car } \alpha_n(-3/2) = \frac{\Gamma(1/2)^2}{2^n n!} \sum_{i=1}^{n-1} \binom{n}{i} \cdot c_i c_{n-i} = \frac{\Gamma(1/2)^2}{2^n n!} \cdot \frac{c_n}{2}.$$

Donc d'après le lemme 20, on a la relation de récurrence sur la suite $(T_n^{-3/2})_{n\geq 1}$: soient $n \geq 2$ et $C \in \mathcal{C}_n$, k et $n-k$ les tailles des deux sous-clades de la racine de C et $C_k \in \mathcal{C}_k$ et $C_{n-k} \in \mathcal{C}_{n-k}$ ces deux sous-clades, on a

$$\begin{aligned} T_n^{-3/2}(C) &= \frac{2 \cdot q_n(k)}{\binom{n}{k}} \cdot T_k^{-3/2}(C_k) \cdot T_{n-k}^{-3/2}(C_{n-k}) \\ &= \frac{c_k c_{n-k}}{c_n} \cdot T_k^{-3/2}(C_k) \cdot T_{n-k}^{-3/2}(C_{n-k}). \end{aligned}$$

Or cette relation est également vérifiée de manière immédiate par la suite $(T_n^{\mathrm{PDA}})_{n\geq 1}$ et donc, par récurrence sur la taille des cladogrammes, $(T_n^{-3/2})_{n\geq 1} = (T_n^{\mathrm{PDA}})_{n\geq 1}$.

- Pour $\beta = -1$, pour $1 \leq i \leq n-1$:

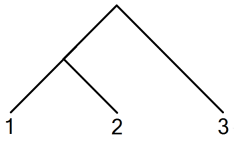
$$q_n^{-1}(i) = \frac{1}{\alpha_n(-1)} \cdot \frac{\Gamma(-1+i+1)\Gamma(-1+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)} = \frac{1}{\alpha_n(-1)} \cdot \frac{1}{i(n-i)}$$

$$\text{et } \alpha_n(-1) = \sum_{i=1}^{n-1} \frac{1}{i(n-i)} = \sum_{i=1}^{n-1} \frac{1}{n}\left(\frac{1}{i} + \frac{1}{n-i}\right) = \frac{2h_{n-1}}{n}.$$

D'où le résultat annoncé dans la proposition. ■

Remarque : Le modèle Bêta vérifie la propriété (i), comme tout modèle de branchements de Markov, mais ne vérifie pas en général la propriété (ii).

Vérifions le dans le cas du modèle AB ($\beta = -1$). On considère C le cladogramme suivant :



Avec le lemme 20 et la proposition 25 on calcule :

$$T_3^{\mathrm{AB}}(C) = \frac{2 \cdot q_3^{\mathrm{AB}}(1)}{\binom{3}{1}} = \frac{1}{3},\ T_5^{\mathrm{AB}}(\varphi_{5,3}^{-1}(C)) = \frac{43}{3 \cdot 25 \cdot 11} \text{ et } T_5^{\mathrm{AB}}(\mathcal{A}_{5,3}) = \frac{32}{25 \cdot 11}.$$

On a donc :

$$\frac{T_5^{\mathrm{AB}}(\varphi_{5,3}^{-1}(C))}{T_5^{\mathrm{AB}}(\mathcal{A}_{5,3})} = \frac{43}{3 \cdot 32} \neq \frac{1}{3} = T_3^{\mathrm{AB}}(C).$$

Mais, comme on le verra en 5, les modèles ERM et PDA, qui vérifient (ii), ne correspondent pas aux données, alors que le modèle AB fournit des cladogrammes proches de ceux des données. La propriété (ii) ne semble donc pas pertinente pour obtenir un modèle correspondant aux données.

5 Mise en évidence du modèle AB

5.1 Taille du plus petit sous-clade

On utilise ici le critère de déséquilibre défini en 4.1 : à chaque nœud, on associe le couple (m, i), où m est la taille du clade associé à ce nœud et i est la taille du plus petit de ses deux sous-clades fils.

Pour un modèle de branchements de Markov donné associé à la suite $(q_n)_{n\geq 2}$, pour $m \geq 2$, on note I_m la taille du plus petit sous-clade d'un clade de m espèces sous ce modèle. La loi de I_m est donnée par :

$$\text{pour } 1 \leq i \leq m/2, \mathbb{P}(I_m = i) = \begin{cases} q_m(i) + q_m(m-i) & \text{si } i \neq m/2 \\ q_m(m/2) & \text{si } i = m/2 \end{cases}.$$

On note alors k_m la médiane de I_m.

Proposition 26

Quand $m \longrightarrow +\infty$, on a les équivalents suivants pour k_m :

- Pour $\beta = -2$, $k_m \sim 1$.
- Pour $\beta = -3/2$, $k_m \sim 1$.
- Pour $\beta = -1$, $k_m \sim \sqrt{m}$.
- Pour $\beta = 0$, $k_m \sim m/4$.
- Pour $\beta = +\infty$, $k_m \sim m/2$.

Démonstration :

- Pour $\beta = -2$, on est sous le modèle du peigne donc $\mathbb{P}(I_m = 1) = 1$. Donc pour tout $m \geq 2$, $k_m = 1$.
- Pour $\beta = -3/2$, on est sous le modèle PDA, donc pour $m \geq 2$:

$$\mathbb{P}(I_m = 1) = 2q_m^{\mathrm{PDA}}(1) = 2\binom{m}{1}\frac{c_1 c_{m-1}}{c_m} = \frac{2m}{2-3} > \frac{1}{2}.$$

  On a donc $k_m = 1$.
- Pour $\beta = -1$, quand $m \longrightarrow +\infty$:

$$\begin{aligned}\mathbb{P}(I_m \leq \sqrt{m}) &= \sum_{1\leq i\leq\sqrt{m}} 2 \cdot \frac{m}{2h_{m-1}} \cdot \frac{1}{i(n-i)} \\ &= \frac{1}{h_{m-1}} \sum_{1\leq i\leq\sqrt{m}} \frac{1}{i} + \frac{1}{n-i} \\ &\sim \frac{1}{\ln(m)} \cdot (\ln(\sqrt{m}) + \frac{\sqrt{m}}{m}) \\ &\sim \frac{1}{2}.\end{aligned}$$

  Et de même $\mathbb{P}(I_m \geq \sqrt{m}) \sim 1/2$, donc $k_m \sim \sqrt{m}$.
- Pour $\beta = 0$,

$$\text{pour } 1 \leq i < m/2, \mathbb{P}(I_m \leq i) = \frac{2i}{m-1} \text{ et donc } k_m = \left\lfloor \frac{m-1}{4} \right\rfloor + 1.$$

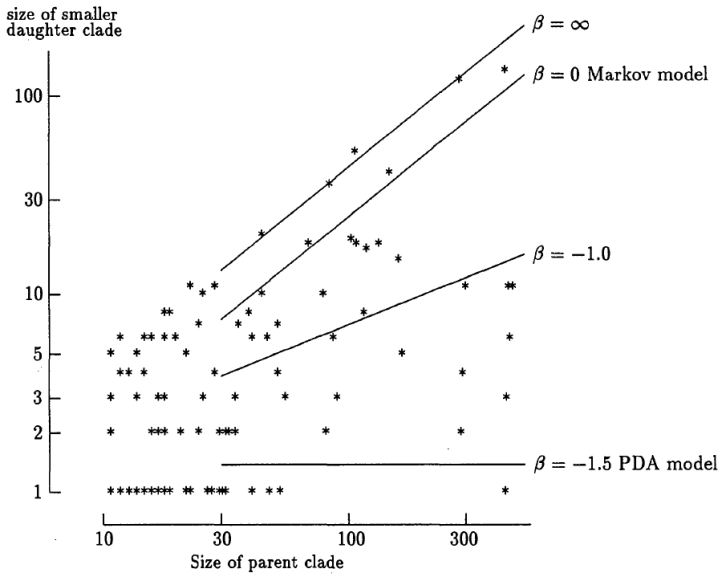
  Donc, quand $m \longrightarrow +\infty$, $k_m \sim m/4$.
- Pour $\beta = +\infty$, on a d'après la remarque du 4.3 pour $1 \leq i < m/2$, $q_m^{+\infty}(i) = \binom{m}{i}/(2^m - 2)$. Alors :

$$\mathbb{P}(I_m \leq \frac{m}{2}) = \frac{1}{2^{m-1}-1}\sum_{i=1}^{m/2}\binom{m}{i} = \begin{cases} 1/2 + \binom{m}{m/2}/(2^{m-1}-1) & \text{si } m \text{ est pair} \\ 1/2 & \text{si } m \text{ est impair} \end{cases}.$$

  Donc $\mathbb{P}(I_m \geq m/2) = \mathbb{P}(I_m \leq m/2) \geq 1/2$ et $k_m = m/2$. ■

Remarque : Dans le cas du modèle PDA ($\beta = -3/2$), quand $m \longrightarrow +\infty$, $\mathbb{P}(I_m = 1) \longrightarrow 1/2$ et $\mathbb{P}(I_m \geq 2) \longrightarrow 1/2$, donc on peut estimer $k_m \sim 3/2$ pour marquer la différence avec le déséquilibre maximal du cas $\beta = -2$, comme le fait Aldous dans [2] et en particulier sur le graphe qui suit.

Dans [2], pour comparer les modèles aux données Aldous part d'un arbre phylogénétique trouvé dans TreeBASE, et place dans un diagramme l'ensemble des couples (m, i) de l'arbre pour $m \geq 10$ (les petites valeurs de m étant peut significatives) avec les valeurs de m en abscisse et celles de i en ordonnée, dans une échelle log-log. Il trace sur le même diagramme les courbes asymptotiques de la médiane k_m pour différentes valeurs de β. Ci-dessous, le diagramme obtenu pour un arbre phylogénétique sur 475 espèces ([2] Figure 3) :



On observe le résultat annoncé en 4.1 : le modèle ERM génère des arbres plus équilibrés que les données et le modèle PDA des arbres plus déséquilibrés. En revanche, le modèle AB (pour $\beta = -1$) semble mieux correspondre aux données.

Cependant cette observation reste assez approximative ici, mais nous allons voir dans les sections suivantes des critères plus précis pour mettre en avant le fait que le modèle AB est très proche des données.

5.2 Maximum de vraisemblance

Dans leur article [4], Blum et François montrent la proximité entre le modèle AB et les données avec des méthodes plus précises : des estimées du maximum de vraisemblance et un critère statistique.

Le modèle β se prête particulièrement bien à une étude des données par des estimées du maximum de vraisemblance car on dispose d'une famille de distributions à un paramètre β qui couvre un large spectre de formes possibles pour les arbres.

En outre, étant donné un arbre $C \in \mathcal{C}_n$, il est facile d'évaluer sa probabilité sous T_n^β : si pour N un nœud interne de C, on note (m_N, i_N) le couple correspondant à la taille du clade associé à N et la taille de son plus petit sous-clade, alors, d'après le lemme 20,

$$T_n^\beta(C) = \prod_{N \text{ nœud de } C} \frac{2 \cdot q_{m_N}^\beta(i_N)}{\binom{m_N}{i_N}}.$$

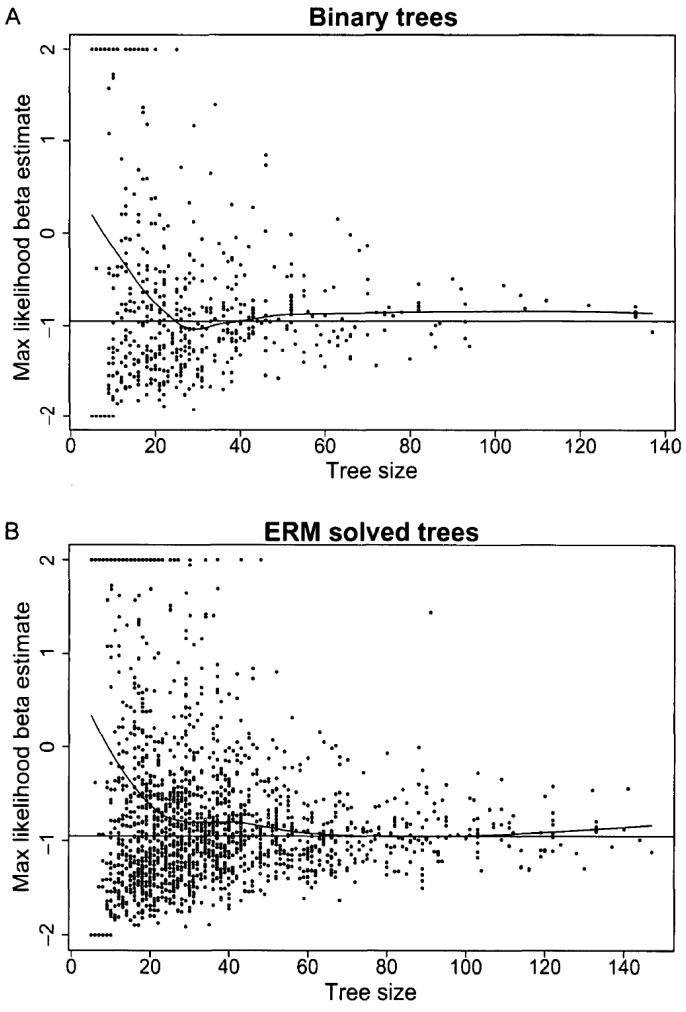
Les $q_n^\beta(i)$ peuvent s'évaluer à partir de la fonction Γ grâce aux formules explicitées dans la remarque du 4.3. On peut alors associer à chaque arbre C une estimée $\hat{\beta}$ de son maximum de vraisemblance : $\hat{\beta}$ est le β qui maximise $T_n^\beta(C)$.

On cherche à calculer $\hat{\beta}$ pour le plus d'arbres possibles dans les données. Malheureusement, tous les arbres trouvés dans les données de TreeBASE ne sont pas *complètement résolus*, c'est-à-dire parfaitement binaires : il y a souvent des polytomies, c'est-à-dire des nœuds ayant strictement plus de 2 descendants. Pour obtenir un arbre binaire sur lequel on peut évaluer $\hat{\beta}$, il faut résoudre chaque polytomie en la remplaçant par des

nœuds binaires choisis aléatoirement, soit sous le modèle ERM, soit sous le modèle PDA. On parle alors d'arbres *ERM-résolus* ou *PDA-résolus*.

Beaucoup d'arbres dans les données ont également dès la racine un sous-clade d'une ou deux espèces très éloignées des autres, qui sert à placer la racine. Ceci conduit à un excès de déséquilibre si on utilise directement ces arbres comme données. Pour éviter cela, les arbres ont été traités par une procédure éliminant automatiquement un sous-clade de la racine qui ne contiendrait qu'une ou deux espèces.

Blum et François ont alors réalisé des estimées du maximum de vraisemblance sur trois groupes d'arbres : des arbres complètement résolus, des arbres ERM-résolus et des arbres PDA-résolus. Pour chaque groupe de données, ils ont représenté les estimateurs $\hat{\beta}$, en fonction de la taille des arbres, la courbe de régression locale de ces valeurs et la droite $\beta = -0.95$. Ils obtiennent les deux graphes suivants, le groupe PDA-résolu n'est pas représenté car il est cohérent avec les deux autres ([4] Figure 2) :



On observe que la courbe de régression locale tend rapidement vers $\beta \approx -0.95$, très proche de la valeur $\beta = -1$ du modèle AB, qui semble donc correspondre de manière assez satisfaisante aux données.

Le tableau suivant ([4] Table 1) regroupe, pour chacun des groupes précédents, les médianes et variances de $\hat{\beta}$ pour 5 intervalles de tailles d'arbres, choisis de manière à ce que chaque intervalle contienne autant d'arbre :

a. Binary trees					
No. taxa	5–11	12–19	20–27	28–43	44–297
Median of $\hat{\beta}$	−.45	−.83	−1.12	−1.02	−0.89
Variance of $\hat{\beta}$	2.97	1.52	.42	.28	.16
b. ERM-solved trees					
No. taxa	5–16	17–24	25–34	35–51	52–536
Median of $\hat{\beta}$	−.74	−.86	−.95	−.94	−.95
Variance of $\hat{\beta}$	2.40	1.20	.52	.39	.12

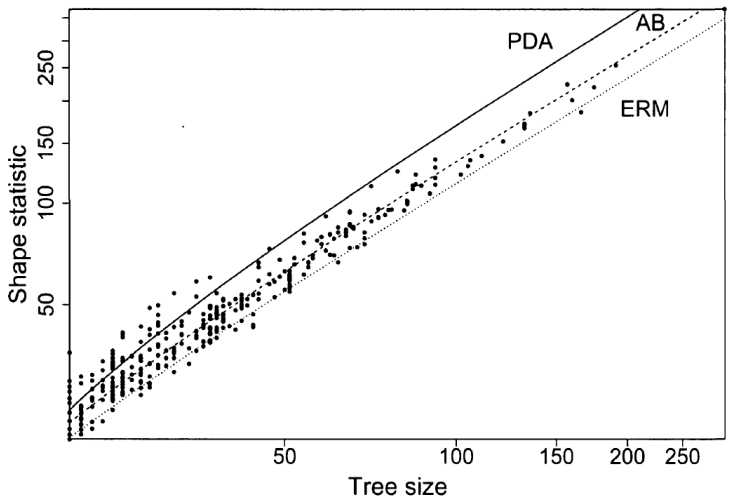
On remarque que les petits arbres sont plus équilibrés ($\hat{\beta}$ plus grand) mais cela n'a pas vraiment de signification car le parti-pris de supprimer les sous-clades de la racine ne contenant qu'une ou deux espèces peut ajouter un équilibre considérable aux petits arbres. On observe également une décroissance rapide de la variance quand la taille des arbres augmente, tandis que la médiane de $\hat{\beta}$ se rapproche de -1.

5.3 Statistique sur la forme des arbres

Pour mesurer l'équilibre d'un arbre, Blum et François introduisent la statistique suivante pour un arbre C, avec les notations précédentes pour un nœud N de C :

$$s = \sum_{N \text{ nœud de } C} \log(m_N - 1).$$

Sur le graphe suivant ([4] Figure 3), les points représentent la valeur de s en fonction de la taille du cladogramme pour des arbres trouvés dans les données. Sont également représentées les prédictions pour les modèles PDA, AB et ERM, calculées à partir de répliques de Monte-Carlo.



On constate de nouveau que le modèle AB est très proche des données. Pour voir cela de manière plus précise, toujours en utilisant la statistique s, on utilise la méthode suivante. On fait l'hypothèse que les données suivent un certain modèle (ici le modèle ERM, PDA ou AB). On pose S une variable aléatoire correspondant à la valeur de s pour un arbre aléatoire suivant le modèle choisi. Pour chaque arbre des données issues

de TreeBASE, on calcule sa valeur de s, et on utilise des répliques de Monte-Carlo pour estimer la P-value $\mathbb{P}(S \geq s)$. Alors si l'hypothèse est vraie, les P-values obtenues ainsi doivent être distribuées uniformément sur $[0, 1]$.

Plus précisément, comme ici les P-values ne prennent qu'un nombre fini de valeurs dans $[0, 1]$ (puisqu'on n'a qu'un nombre fini d'arbre dans les données), on a : si l'hypothèse est vraie, pour toute valeur v atteinte par une P-value, la proportion de P-values inférieures ou égales à v est exactement v. Ceci est une conséquence du lemme suivant :

Lemme 27

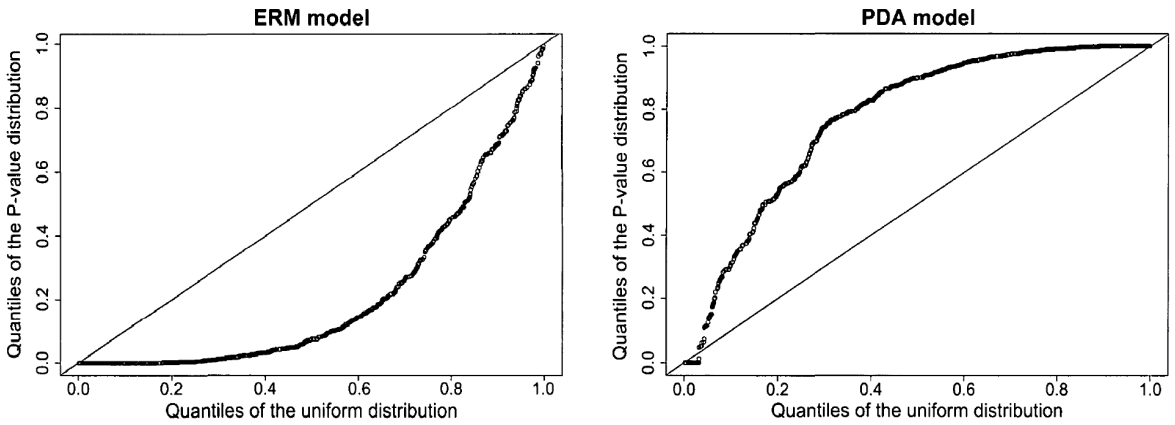
Soit X une variable aléatoire réelles à valeurs dans un ensemble fini E. Pour $x \in \mathbb{R}$, on note $G(x) = \mathbb{P}(X \geq x)$. La fonction G est à valeurs dans un ensemble fini F. Alors pour tout $y \in F$, $\mathbb{P}(G(X) \leq y) = y$.

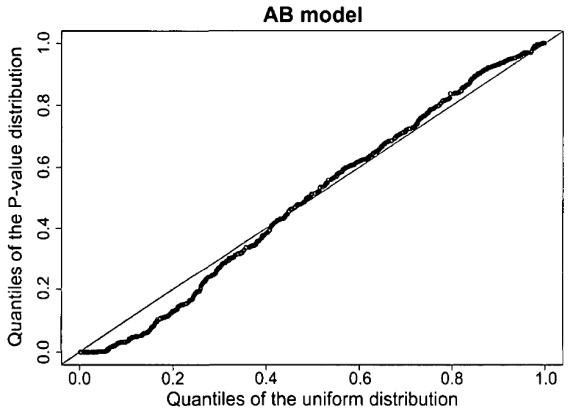
Démonstration :

- Notons P_X la loi de X et E' l'ensemble des points de E de masse non nulle sous P_X. Pour $x \in \mathbb{R}$,
$$G(x) = \mathbb{P}(X \geq x) = P_X([x, +\infty[) = \sum_{\substack{a \in E' \\ a \geq x}} P_X(a).$$
G est donc une fonction en escalier décroissante continue à gauche et dont l'ensemble des points de discontinuités est E'.
- Soit $y \in F$. D'après les observations précédentes sur la fonction G, $G^{-1}(\{y\})$ est un intervalle de la forme $]b, c]$ avec $b, c \in E'$ consécutifs $(]b, c[\cap E' = \varnothing$. Comme G est décroissante, $G^{-1}(]-\infty, y]) =]b, +\infty[$ et donc :
$$\mathbb{P}(G(X) \leq y) = P_X(]b, +\infty[) = \sum_{\substack{a \in E' \\ a > b}} P_X(a) = \sum_{\substack{a \in E' \\ a \geq c}} P_X(a) = G(c) = y.$$

■

Pour un modèle donné, on représente, pour chaque arbre des données de valeur s, la proportion de P-values inférieures ou égales à $\mathbb{P}(S \geq s)$ en fonction de la proportion d'arbres des données ayant une valeur $s' \geq s$. Si les arbres des données suivent bien le modèle, l'ensemble des points tracés doit être inclus dans la diagonale du graphe. Dans le cas des modèles ERM, PDA et AB, on a les graphes suivants ([4] Figure 4) :





Ces graphes confirment que les arbres des données ont une distribution assez proche de celle fournie par le modèle AB.

5.4 Processus d'évolution sous-jacent

Le modèle AB semble correspondre aux données, mais il est uniquement issu d'une définition mathématique. On se demande alors si l'on ne peut pas l'obtenir à partir d'un modèle d'évolution des espèces au fil du temps (c'est-à-dire qu'il vérifierait la propriété (b)). Dans leur article [4], Blum et François proposent un modèle à un paramètre construit à partir de (b), assez proche du modèle Bêta et donc en particulier du modèle AB.

On a vu qu'on ne peut pas obtenir un tel modèle à partir de la classe de modèles définie en 2.2 puisque les cladogrammes aléatoires fournis par ces modèles sont ceux du modèle ERM. La particularité de ces modèles était le fait que lors d'une spéciation ou d'une extinction, toutes les espèces avaient la même probabilité d'être l'espèce concernée par l'évènement. Le modèle introduit par Blum et François fait donc intervenir des variations entre les taux de spéciations des différentes espèces (et ne fait pas intervenir d'extinction).

On part d'une espèce avec un taux de spéciation λ_0. On suppose qu'à chaque spéciation d'une espèce ayant un taux de spéciation λ, l'une des espèces filles obtient un taux de spéciation λp et l'autre le taux $\lambda(1-p)$ avec $p \in [0,1]$ (le choix de l'espèce ayant le taux λp étant uniforme parmi les deux espèces) et elles conservent ensuite ces taux jusqu'à se séparer à leur tour. Pour $n \geq 1$ on arête le processus après $n-1$ spéciations, on obtient un cladogramme aléatoire à n feuilles en distribuant de manière uniforme les étiquettes de 1 à n sur les espèces.

De manière analogue à celle utilisée pour décrire la classe de modèles de 2.2, on peut noter τ_n l'instant de la $n^{\text{ème}}$ spéciation et X_n le cladogramme aléatoire à n feuilles obtenu entre τ_{n-1} et τ_n. Si l'espèce i de X_n a pour taux de spéciation μ_i, alors on a $\tau_n - \tau_{n-1} = \min(\zeta_1, \ldots, \zeta_n)$ où $\zeta_1, \ldots, \zeta_n$ sont des variables aléatoires exponentielles indépendantes de paramètres $\mu_1, \ldots, \mu_n$ et l'espèce qui spécie est l'espèce i telle que $\zeta_i = \tau_n - \tau_{n-1}$ (i est unique p.s.). On note que $\mu_1 + \cdots + \mu_n = \lambda_0$ est un invariant du processus.

Montrons tout d'abord quelques résultats élémentaires concernant les variables aléatoires exponentielles.

Lemme 28

Soient $\zeta_1, \ldots, \zeta_n$ des variables exponentielles indépendantes de paramètres $\mu_1, \ldots, \mu_n$. On pose $\xi = \min(\zeta_1, \ldots, \zeta_n)$. Alors ξ est une variable aléatoire exponentielle de paramètre $\mu_1 + \cdots + \mu_n$ et pour $1 \leq i \leq n$,

$$\mathbb{P}(\zeta_i = \xi) = \frac{\mu_i}{\mu_1 + \cdots + \mu_n}.$$

Démonstration :

- Soit $t \geq 0$,

$$\mathbb{P}(\xi > t) = \mathbb{P}(\zeta_1 > t, \ldots, \zeta_n > t) = \mathbb{P}(\zeta_1 > t) \cdots \mathbb{P}(\zeta_n > t) = e^{-\mu_1 t} \cdots e^{-\mu_n t}.$$

  On en déduit $\mathbb{P}(\xi \leq t) = 1 - e^{-(\mu_1 + \cdots + \mu_n)t}$ donc ξ a la fonction de répartition d'une variable aléatoire de paramètre $\mu_1 + \cdots + \mu_n$.

- Soit $1 \leq i \leq n$.
  D'après le 1er point, la variable aléatoire $\xi_i = \min(\zeta_1, \ldots, \zeta_{i-1}, \zeta_{i+1}, \ldots, \zeta_n)$ est exponentielle de paramètre $\rho_i = \mu_1 + \cdots + \mu_{i-1} + \mu_{i+1} + \cdots + \mu_n$ et est indépendante de ζ_i. On a donc

$$\begin{aligned}
\mathbb{P}(\zeta_i = \xi) &= \mathbb{P}(\zeta_i \leq \xi_i) \\
&= \int_{(\mathbb{R}_+)^2} \mathbb{1}_{x \leq y}\ \mu_i e^{-\mu_i x} \rho_i e^{-\rho_i y} dx dy \\
&= \int_0^{+\infty} \left(\int_x^{+\infty} \rho_i e^{-\rho_i y} dy \right) \mu_i e^{-\mu_i x} dx \\
&= \int_0^{+\infty} e^{-\rho_i x} \mu_i e^{-\mu_i x} dx \\
&= \frac{\mu_i}{\mu_1 + \cdots + \mu_n}.
\end{aligned}$$

■

Ce résultat signifie que si les espèces de X_n ont pour taux de spéciation $\mu_1, \ldots, \mu_n$, alors la probabilité que ce soit l'espèce i qui spécie à l'instant τ_n est $\mu_i/(\mu_1 + \cdots + \mu_n)$.

Lemme 29

Le modèle défini ci-dessus est le modèle de branchements de Markov associé à la suite $(q_n)_{n \geq 2}$ où pour $1 \leq i \leq n-1$

$$q_n(i) = \frac{1}{2} \cdot \binom{n-2}{i-1} (p^{i-1}(1-p)^{n-i-1} + (1-p)^{i-1} p^{n-i-1}).$$

Démonstration : On suppose que lors du procédé, lorsqu'une espèce spécie, les deux espèces filles obtiennent des marqueurs gauche-droite. Ces marqueurs sont ensuite enlevés pour obtenir le cladogramme associé au procédé.

On note $G = \{$l'espèce gauche de la racine a pour taux de spéciation $p\lambda_0\}$ et D son complémentaire. Pour $1 \leq i \leq n-1$, on note I_n la taille du sous-clade gauche de la racine entre les instants τ_{n-1} et τ_n.

- Montrons par récurrence sur $n \geq 2$ que pour tout $1 \leq i \leq n-1$,

$$\mathbb{P}(I_n = i \mid G) = \binom{n-2}{i-1} p^{i-1}(1-p)^{n-i-1}.$$

C'est évidemment vrai pour $n = 2$. Soit $n \geq 2$, supposons le résultat vrai au rang n. Si $I_{n+1} = i$, alors $I_n = i$ ou $I_n = i-1$ et ces évènements sont disjoints, donc :

$$\begin{aligned}\mathbb{P}(I_{n+1} = i \mid G) = \mathbb{P}(I_{n+1} = i \mid I_n = i \text{ et } G)\ \mathbb{P}(I_n = i \mid G) \\ + \mathbb{P}(I_{n+1} = i \mid I_n = i-1 \text{ et } G)\ \mathbb{P}(I_n = i-1 \mid G).\end{aligned}$$

Sachant $I_n = i$ et G, la probabilité que $I_{n+1} = i$ est la probabilité que l'espèce qui spécie à l'instant τ_n appartient au clade droit. D'après le lemme 28, cette probabilité est donc égale au quotient de la somme des taux de spéciation du clade droit (qui vaut $\lambda_0(1-p)$) par la somme des taux de spéciation de toutes les espèces (qui vaut λ_0). On a donc $\mathbb{P}(I_{n+1} = i \mid I_n = i \text{ et } G) = (1-p)$.
De même, sachant $I_n = i-1$ et G, la probabilité que $I_{n+1} = i$ est égale au quotient de la somme des taux de spéciation du clade gauche (qui vaut $\lambda_0 p$) par la somme des taux de spéciation de toutes les espèces (qui vaut λ_0), donc $\mathbb{P}(I_{n+1} = i \mid I_n = i-1 \text{ et } G) = p$.
En appliquant l'hypothèse de récurrence, on obtient

$$\begin{aligned}\mathbb{P}(I_{n+1} = i \mid G) &= (1-p) \cdot \binom{n-2}{i-1} p^{i-1}(1-p)^{n-i-1} + p \cdot \binom{n-2}{i-2} p^{i-2}(1-p)^{n-i} \\ &= \binom{n-1}{i} p^{i-1}(1-p)^{(n+1)-i-1},\end{aligned}$$

ce qui conclut la récurrence.

- Soit $1 \leq i \leq n-1$, l'évènement D correspond au fait que l'espèce gauche de la racine ait le taux de spéciation $\lambda_0(1-p)$ donc en appliquant le 1^{er} point, on a

$$\mathbb{P}(I_{n+1} = i \mid D) = \binom{n-1}{i} (1-p)^{i-1} p^{(n+1)-i-1}.$$

On en déduit

$$\begin{aligned}\mathbb{P}(I_n = i) &= \frac{1}{2}\mathbb{P}(I_n = i \mid G) + \frac{1}{2}\mathbb{P}(I_n = i \mid D) \\ &= \frac{1}{2} \cdot \binom{n-2}{i-1} (p^{i-1}(1-p)^{n-i-1} + (1-p)^{i-1} p^{n-i-1}).\end{aligned}$$

On remarque donc que la probabilité que le clade gauche de la racine ait i espèces entre τ_{n-1} et τ_n ne dépend pas de λ_0.
Or les deux sous-clades de la racine suivent le même modèle, au taux de spéciation initial λ_0 près. Donc on peut appliquer récursivement le résultat précédent à toutes les clades de X_n.
Soit $C \in \mathcal{C}_n$, on note $\mathcal{B}(C)$ l'ensemble des cladogrammes de $\mathcal{C}_n$ qui ont la même forme que C. Soit $(q_n)_{n\leq 2}$ la suite définie dans l'énoncé du lemme et $(T_n)_{n\leq 1}$ la suite de distribution du modèle de branchements de Markov associé à $(q_n)_{n\leq 2}$.
Le poids de $\mathcal{B}(C)$ sous la loi de X_n s'obtient uniquement à partir des probabilités de branchements $q_k(i)$ (on peut obtenir la valeur précise par un raisonnement analogue

à celui du lemme 20 en ne tenant pas compte des étiquettes) et donc il se calcule de la même manière que $T_n(\mathcal{B}(C))$. On a donc

$$\mathbb{P}(X_n \in \mathcal{B}(C)) = T_n(\mathcal{B}(C)) \text{ et } \mathbb{P}(X_n \in C) = \frac{\mathbb{P}(X_n \in \mathcal{B}(C))}{\#\mathcal{B}(C)} = \frac{T_n(\mathcal{B}(C))}{\#\mathcal{B}(C)} = T_n(C)$$

car la loi de X_n et T_n vérifient la propriété (i) donc le poids qu'elles donnent a un arbre de dépend que de sa forme, donc tous les éléments de $\mathcal{B}(C)$ ont même poids. ■

Pour définir le *modèle BB* (pour *Bêta-Binomial*), Blum et François rajoutent alors un second niveau d'aléatoire en supposant que p n'est pas déterministe, mais qu'il suit la loi $\text{Beta}(\alpha+1, \alpha+1)$ pour $\alpha > -1$.

Pour chaque $\alpha > -1$, le modèle BB de paramètre α est le modèle de branchements de Markov associé à la suite $(q_n^{\text{BB},\alpha})_{n\geq 2}$ obtenue en moyennant sur p : pour $1 \leq i \leq n-1$

$$\begin{aligned} q_n^{\text{BB},\alpha}(i) &= \mathbb{E}\left[\frac{1}{2} \cdot \binom{n-2}{i-1}(p^{i-1}(1-p)^{n-i-1} + (1-p)^{i-1}p^{n-i-1})\right] \\ &= \frac{1}{2} \cdot \binom{n-2}{i-1} \int_0^1 (x^{i-1}(1-x)^{n-i-1} + (1-x)^{i-1}x^{n-i-1})x^\alpha(1-x)^\alpha dx \\ &= \frac{1}{2} \cdot \binom{n-2}{i-1}(\text{B}(i+\alpha, n-i+\alpha) + \text{B}(n-i+\alpha, i+\alpha)) \\ &= \frac{1}{b_n(\alpha)} \frac{\Gamma(i+\alpha)\Gamma(n-i+\alpha)}{\Gamma(i)\Gamma(n-i)} \end{aligned}$$

avec la constante normalisatrice

$$b_n(\alpha) = \frac{\Gamma(n+2\alpha)}{\Gamma(n-1)}.$$

En comparant cette formule à celle obtenue en 4.3 pour le modèle Bêta

$$q_n^\beta(i) = \frac{1}{\alpha_n(\beta)} \cdot \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)},$$

on constate qu'elles sont très proches. Les deux reposent sur une séparation binomiale dont le paramètre p suit une loi Beta. Mais une distribution binomiale $\text{bin}(n,p)$ donne une mesure non nulle aux entiers de 0 à n, alors qu'une distribution de séparation q_n porte sur les entiers de 1 à $n-1$.

Comme le montre le lemme 29, le modèle BB repose sur une distribution $\text{bin}(n-2,p)$ translatée de 1 pour qu'elle porte sur les entiers de 1 à $n-1$. Cette distribution est ensuite symétrisée pour obtenir la formule du lemme 29. On a donc dans ce cas un facteur $\binom{n-2}{i-1}$ dans $q_n(i)$, c'est-à-dire un $\Gamma(i)\Gamma(n-i)$ au dénominateur. Le modèle Bêta repose quand à lui sur une distribution $\text{bin}(n,p)$ conditionnellement au fait que les valeurs extrêmes 0 et n ne sont pas atteintes, ce qui mène à un facteur $\binom{n}{i}$ dans $q_n(i)$, c'est-à-dire un $\Gamma(i+1)\Gamma(n-i+1)$ au dénominateur.

À cette différence près, le modèle BB est très proche du modèle Bêta : pour $\alpha = 0$, on retrouve le modèle ERM ($\beta = 0$) et pour $\alpha = -1$, on retrouve le modèle du peigne ($\beta = -2$). En outre, le modèle AB ($\beta = -1$) dont on cherchait un processus d'évolution sous-jacent est très bien approché par le modèle BB pour certaines valeurs de α ([4]).

Conclusion

Les propriétés (i) et (ii) définies en 1.2 nous amènent à construire les modèles PDA et ERM, ainsi que le modèle du peigne, correspondant à un déséquilibre maximal. Mais il semble que l'on ne puisse pas obtenir une famille de distributions à un paramètre caractérisant ces propriétés. On peut en revanche trouver une famille qui contient les 3 modèles précédents : le modèle Bêta.

Le modèle ERM fournit des arbres plus équilibrés que ceux que l'on trouve dans la base de données TreeBASE et le modèle PDA des arbres plus déséquilibrés, mais le modèle AB, obtenu en prenant un paramètre β intermédiaire entre ceux des modèles ERM et PDA, est assez proche des données.

Ce modèle n'a été construit qu'à partir de considérations mathématiques, mais on peut trouver une description temporelle de l'évolution qui amène à une famille de modèles proche de la famille du modèle Bêta. Même si cette description n'est pas simple, elle met en avant le fait que pour obtenir des arbres qui correspondent aux données, il faut que, lors d'une spéciation, les deux espèces filles n'aient pas le même taux de spéciation.

Références

[1] David J. Aldous. Probability distributions on cladograms. In D. J. Aldous and R. Pemantle, editors, *Random Discrete Structures*, pages 1–18. Springer Berlin, 1995.

[2] David J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees from yule to today. *Statistical Science*, 16(1) :23–34, 2001.

[3] Emil Artin. *The Gamma Function.*

[4] Michael G. B. Blum and Olivier Francois. Which random processes describe the tree of life ? a large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4) :685–691, 2006.

[5] Joseph Felsenstein. Numerical methods for inferring evolutionary trees. *The Quartely Review of Biology*, 57(4) :379–404, 1982.

[6] Joseph Felsenstein. Phylogenies from molecular sequences : Inference and reliability. *Annual Reviews of Genetic*, 22 :521–565, 1988.

[7] G. U. Yule. A mathematical theory of evolution, based on the conclusion of dr. j. c. willis. *Philos. Trans. Roy. Soc. London Ser. B*, 213 :21–87, 1924.