UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS
Y EN SISTEMAS

MODELOS MATEMÁTICOS PARA LA GENERACIÓN DE ÁRBOLES FILOGENÉTICOS

TESINA

QUE PARA OBTENER LA ESPECIALIDAD DE:
ESPECIALISTA EN ESTADÍSTICA APLICADA

PRESENTA:
KARINA YAÑEZ AROCHE

TUTOR: DR. ARNO SIRI-JÉGOUSSE
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS
Y EN SISTEMAS

CIUDAD DE MÉXICO, ENERO, 2022

Índice

1. Introducción

Este trabajo está basado en el artículo *Ranked Tree Shapes, Nonrandom Extinctions, and the Loss of Phylogenetic Diversity* de Maliet et al. (2018). Mi objetivo es presentar distintos modelos matemáticos que predicen patrones de diversidad de especies y la generación de árboles filogenéticos, así como un resumen de sus propiedades matemáticas. Haré énfasis en la inferencia de parámetros del modelo propuesto por Maliet et al. (2018) quienes investigan cómo la pérdida de diversidad filogenética está influenciada por la forma clasificada del árbol de especies, caracterizada por la relación entre la riqueza de clados y su edad relativa y las extinciones no aleatorias, caracterizadas por la relación entre la riqueza de clados y su abundancia relativa.

Las especies difieren sustancialmente en la cantidad de información genética única con la que cuentan (Faith 1992; Reeding y Mooers 2006). Se han desarrollado varios sistemas de medición para captar la variación genética entre especies, sin embargo, si se les clasificara con fines de conservación basándonos únicamente en la medición de su variación genética, se daría preferencia a las especies que cuentan con cantidades desproporcionadamente grandes de información genética única, por encima de las que tienen muchos parientes cercanos (menor variacion genética intrínseca).

En general, las especies taxonómicamente distintas contribuyen más a la diversidad de un subconjunto dado, porque aportan características diferentes. Dado que estas características no se enumeran explícitamente, la diversidad taxonómica vuelve a plantear la dificultad de evaluar o medir los atributos de interés, y sugiere que se necesita algún indicador medible. Entonces, se introduce la medida de la diversidad filogenética que es un indicador eficaz de la diversidad de las características subyacentes. Con base en Faith (1992), los autores definen esta diversidad filogenética como una medida del legado evolutivo de un grupo de especies, que puede utilizarse para definir las prioridades de conservación, definida también como la suma de las longitudes de las ramas de la filogenia abarcadas por un determinado conjunto de taxones.

En un supuesto evento de extinción, supongamos que k especies se salvan de un total de n. Esto puede hacerse de muchas maneras. En un extremo, las especies pueden elegirse al azar con respecto a sus relaciones filogenéticas; en otro extremo, útil para la comparación, la especie puede elegirse de acuerdo con un algoritmo que maximiza la cantidad de historia evolutiva conservada, usando el coalescente de Kingman (del cual se va a hablar más adelante). Se seleccionan los nodos $k-1$ más bajos de un árbol (contando desde la raíz). Esto define a k-clados. Una especie de cada clado es seleccionada, si un clado tiene más de una especie en él, entonces se escoge una al azar. Este algoritmo optimiza la cantidad de historia evolutiva conservada en relación con la pérdida de especies. Si se salvan k especies de un total de n, es natural expresar la cantidad de historia conservada como una fracción de la cantidad total que podría haberse conservado si se hubieran salvado todas las n especies (Nee y May 1997).

Con respecto a lo anterior, Nee y May (1997) encontraron que el 80 % de la diversidad filogenética puede ser conservada incluso cuando se pierda el 95 % de las especies, esta pérdida

filogenética es mayor cuando se utilizan otros modelos de diversificación de especies, tales como los modelos de Yule y nacimiento-muerte; estos modelos generan bordes colgantes más largos (es decir, ramas que conducen a las puntas), por lo que la redundancia filogenética es menor que en el coalescente de Kingman. Esta redundancia filogenética se refiere a la reducción de diversidad que se obtiene al tomar en cuenta relaciones evolutivas entre especies.

Maliet et al. (2018) esperan que la correlacion entre la riqueza de especies del clado y su edad relativa tengan impacto en la perdida de diversidad filogenética. La edad relativa, se refiere a la profundidad (medida desde el presente) de su nodo raíz (es decir, el nodo donde el clado está atado al resto del árbol). En caso de extinción aleatoria, dado que los clados más pequeños tienen más probabilidades de extinguirse primero, la consecuencia de su extinción total en la diversidad filogenética dependerá de la longitud de los bordes colgantes de estos clados en comparación con los de los clados más grandes.

Por otro lado, se ha demostrado que la pérdida filogenética también se ve influenciada por la distribución de los riesgos de extinción dentro del árbol de especies (árbol de la vida), donde escenarios más realistas predicen pérdidas más grandes de diversidad filogenética que escenarios con riesgos de extinción aleatorias. Se ha sugerido que la tasa de pérdida de la diversidad filogenética en el futuro podría ser mucho mayor que la tasa de pérdida de especies porque las especies amenazadas no se distribuyen al azar en la filogenia (Vernon et al. 2015). En el caso del modelo de field of bullets, se consideran riesgos de extinción con distribuciones iguales en todas las especies, sin embargo, se puede asumir que los eventos de extinción son independientes, pero no igualmente distribuídos. Una extensión más realista permite a cada especie tener su propia probabilidad de supervivencia, bajo este modelo, buscamos predecir la puntuación de diversidad filogenética del conjunto de taxones que sobreviven. Esta diversidad filogenética futura es una variable aleatoria con una distribución bien definida, pero hasta la fecha, la mayor parte de la atención se ha centrado sólo en su media, es decir, en la puntuación de diversidad filogenética esperada de las especies que sobreviven (Faller et al. 2008; Vernon 2015).

2. Propiedades de árboles filogenéticos

En particular, hay dos características deseables e importantes en algunos modelos de generación de árboles aleatorios (por ejemplo, modelo de Yule, Alpha-splitting o Beta-splitting). Estas dos carácterísticas son la intercambiabilidad y consistencia de muestreo.

2.1. Intercambiabilidad

La intercambiabilidad se refiere al hecho de que volver a etiquetar las hojas o puntas del árbol, este no cambia su probabilidad. La intercambiabilidad requiere que la probabilidad de un árbol bajo una distribución particular dependa sólo de la forma del árbol. Por lo tanto, solo necesitamos considerar las distribuciones en el conjunto de formas de árboles. Se trata

de una condición natural, ya que no permite que los nombres o etiquetas de las especies jueguen un papel especial en la distribución de probabilidad (Hollering and Sullivant 2019; Maliet et al. 2018).

2.2. Consistencia de muestreo

La consistencia de muestreo implica que al remover una punta etiquetada $n+1$ (o cualquier otra por intercambiabilidad) de un cladograma aleatorio con $n+1$ puntas, no modifica la forma del árbol clasificado que se obtiene con n puntas pues la distribución de estas puntas continua siendo la misma (Ford 2005; Maliet et al. 2018). La consistencia del muestreo es una condición natural para un modelo de árbol aleatorio porque significa que las especies que faltan al azar no afectan la distribución subyacente de las especies que se observaron.

2.3. Topología, forma y balance del árbol

La topología de árboles se refiere a los patrones de relación evolutiva entre un grupo de especies, esto es independiente de las longitudes de las ramas de un árbol filogenético, es decir, es independiente de la edad de las especies. A veces dos árboles pueden verse muy diferentes, pero tienen la misma topología (Figura 1).

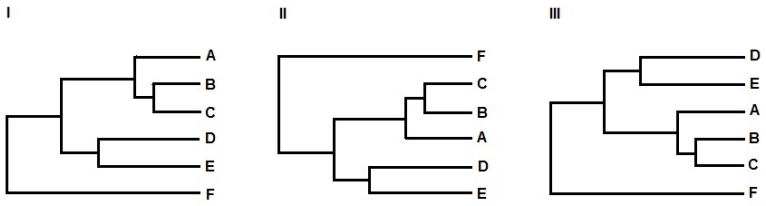

Figura 1: Tres árboles filogenéticos que muestran diferentes formas de trazar la misma topología.

La forma del árbol ignora tanto la longitud de las ramas como las etiquetas de las puntas del árbol (nombre de las especies). En la figura 2, I y II los nodos tienen los mismos patrones en términos del número de descendientes en cada lado de la bifurcación, mientras III tiene una forma diferente.

Un aspecto de la forma del árbol es el equilibrio o balance del árbol (Figura 3). El balance generalmente se refiere a la estructura topológica del árbol, nuevamente, sin considerar las longitudes de las ramas. El balance es una forma de expresar diferencias en el número de descendientes entre pares de linajes en diferentes puntos de un árbol filogenético. Muchos trabajos se centran en el balance de un árbol, el cual puede ser medido por diversos índices,

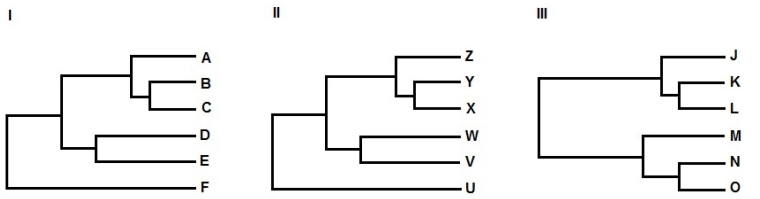

Figura 2: I y II son árboles filogenéticos que comparten la misma forma, y III con una forma diferente.

por ejemplo, índice de Colless e índice de Sackin (Heard 1992; Sainudiin and Véber 2016). Diferentes modelos de generación de árboles pueden dar como resultado árboles con diferentes grados de balance sin importar qué índice se utilice. El balance de los árboles es una consideración importante para la sistemática filogenética, porque afectará la precisión de sus estimaciones.

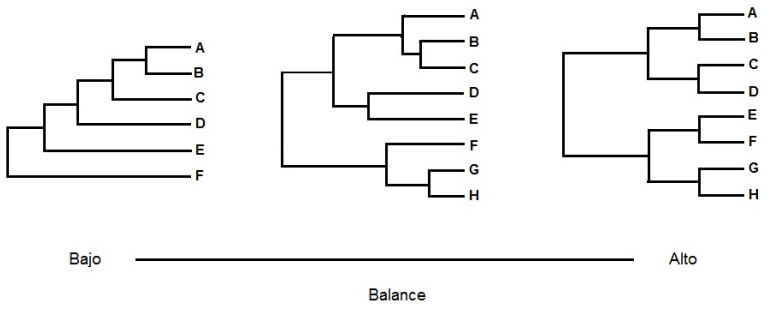

Figura 3: El balance se refiere a la distribución de taxones entre los diferentes clados de una filogenia. Si los taxones se distribuyen uniformemente entre los clados, entonces la topología está balanceada.

Un modelo que nos permite ejemplificar lo anterior lo podemos encontrar en Aldous (1996; 2001), quien introduce una familia de cladogramas aleatorios de un parámetro, denominada modelo Beta-splitting, que además explicaremos a detalle más abajo. Aquí, un cladograma se define como una forma de árbol binario con un número específico de puntas (n hojas). Las hojas están etiquetadas por la especie muestreada o por $\{1, ..., n\}$. El parámetro $\beta > -2$ modula la forma y el balance del árbol generado por este modelo al determinar la distribución dividida de un nodo del que colgaran m hojas. En este caso, para β pequeños (cercanos a -1), los árboles correspondientes tendrán una alta probabilidad de ser desequilibrados o desbalanceados, mientras que para β grandes la distribución de árboles se concentra en árboles equilibradoso balanceados, cubriendo así una amplia gama de posibles topologías (Blum and Francois 2006; Sainudiin and Véber 2016).

3. Algunos modelos matemáticos para la generación de árboles filogenéticos

La información filogenética es fundamental para las inferencias sobre macroevolución ya que estamos interesados en el destino de los clados. El modelado matemático del proceso dinámico de especiación y extinción, en general modelos de "nacimiento-muerte", se puede utilizar para responder muchas preguntas en macroevolución.

3.1. Modelo de nacimiento-muerte

Un modelo de nacimiento-muerte es un modelo estocástico de evolución en tiempo continuo en el que el estado, representado por todas las especies que viven en un momento dado, puede cambiar de dos formas posibles:

- evento de especiación (que representa el nacimiento), cuando una especie se divide en exactamente dos especies descendientes, y
- evento de extinción (que representa la muerte), cuando una especie muere.

El proceso evolutivo comienza con una sola especie en algún momento $t_0 > 0$ en el pasado. La tasa de nacimientos y muertes en un momento dado depende de la cantidad de especies existentes. Cuando hay n especies, se produce un nacimiento con una tasa λ_n y una muerte con tasa μ_n.

Ahora consideremos el caso donde en cada linaje el tiempo de espera para el próximo evento de especiación es exponencial con el parámetro λ, y el tiempo de espera para el próximo evento de extinción es exponencial con el parámetro μ, independientemente de los demás linajes. En cada linaje, el tiempo de espera para el siguiente evento es exponencial con el parámetro $\lambda + \mu$, y la probabilidad de que ese evento sea de especiación es

$$\frac{\lambda}{\mu + \lambda} \tag{1}$$

y de extinción

$$\frac{\mu}{\mu + \lambda}.$$

En general, si hay $N(t)$ linajes vivos en el tiempo t, entonces el tiempo de espera hasta el próximo evento sigue una distribución exponencial con el parámetro $N(t)(\lambda + \mu)$, con la probabilidad de que ese evento sea especiación o extinción. Además, los tiempos de espera en todos los linajes se acortan cada vez más a medida que se acumulan más linajes. Es posible definir dos parámetros adicionales, la tasa de diversificación neta (r) y la tasa de extinción relativa (ϵ):

$$r = \lambda - \mu, \qquad \epsilon = \frac{\mu}{\lambda}. \tag{2}$$

La probabilidad de especiación y extinción durante un intervalo de tiempo pequeño, Δt (intervalo tan pequeño que a lo mucho contenga un solo evento, ya sea especiación o extinción, o ninguno), se puede expresar como:

$$Pr_{especiacion} = N(t)\lambda\Delta t \qquad Pr_{extincion} = N(t)\mu\Delta t \tag{3}$$

El valor esperado de $N(t)$ después de Δt es:

$$N(t + \Delta t) = N(t) + N(t)\lambda\Delta t - N(t)\mu\Delta t$$

Podemos convertir lo anterior en una ecuación diferencial restando $N(t)$ de ambos lados, luego dividiendo por Δt y tomando el límite cuando Δt se vuelve muy pequeño:

$$\frac{dN}{dt} = rN(dt). \tag{4}$$

Si establecemos la condición de que $N(0) = n_0$; es decir, en el tiempo 0, comenzamos con n_0 linajes. Entonces obtenemos:

$$N(t) = n_0 e^{rt}. \tag{5}$$

Esta ecuación nos da el valor esperado del número de especies a través del tiempo bajo un modelo de nacimiento-muerte. El número de especies crece exponencialmente a lo largo del tiempo siempre que $\lambda > \mu$, es decir, $r > 0$, y decae en caso contrario.

En el caso mas general donde las tasas no son generales, el mecanismo de evolución está descrito por

$$Pr(N(t, t + \Delta t) - N(t) = 0 | N(t) = n) = 1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t). \tag{6}$$

Para simular árboles filogenéticos a través del tiempo se utilizan las propiedades anteriores del modelo de nacimiento-muerte. El árbol filogenético a menudo comienza con el primer evento de especiación en el clado.

Un posible algoritmo de simulación es el siguiente: Suponemos que tenemos un cierto número de linajes vivos en el árbol (1 o 2 inicialmente), un tiempo actual ($t_i = 0$) y un tiempo de

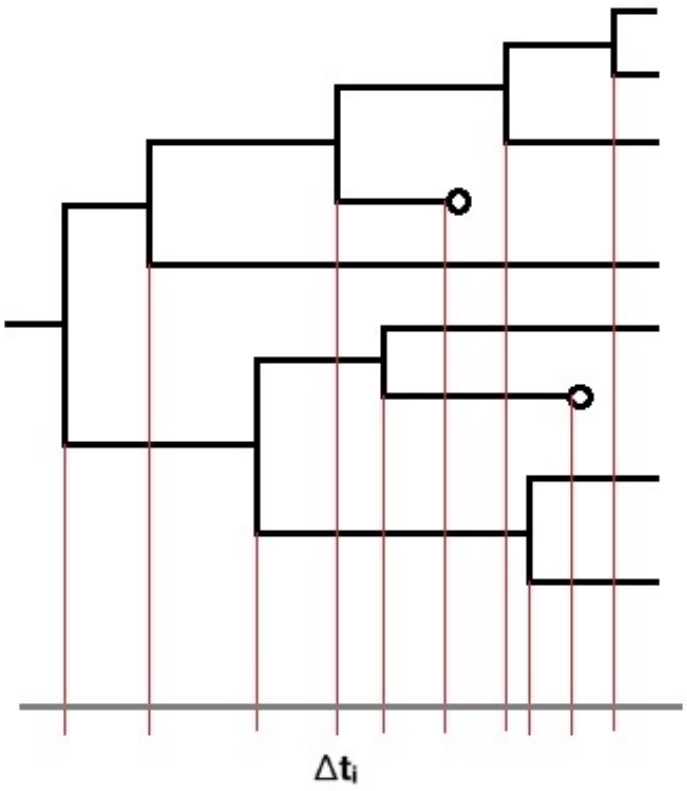

Figura 4: Un arbol de nacimiento-muerte con tiempos de espera Δt, donde los circulos denotan taxones extintos.

parada t_{stop}.

1. Se dibuja un tiempo de espera hasta el próximo evento de especiación o extinción t_{i+1}. Los tiempos de espera se extraen de una distribución exponencial con el parámetro de tasa $\lambda_{N_{t_i}} + \mu_{N_{t_i}}$, donde N_{t_i} es el número actual de linajes vivos en el árbol.

2. Si la simulación termina antes del próximo evento. Es decir, si $t_{i+1} > t_{stop}$, finaliza la simulación.

3. Después se decide si el próximo evento es un evento de especiación, con probabilidad $\frac{\lambda}{\mu+\lambda}$ o un evento de extinción, con probabilidad $\frac{\mu}{\mu+\lambda}$. Esto se puede hacer dibujando un número aleatorio uniforme u_i del intervalo $(0, 1)$ y asignando especiación al evento si $u_i < \frac{\lambda}{\mu+\lambda}$ y extinción en caso contrario.

4. Si el evento en el paso 3 es un evento de especiación, se elije un linaje vivo aleatorio

en el árbol. Se adjunta una nueva rama al árbol en este punto y se agrega un nuevo linaje vivo a la simulación. Se regresa al paso 1.

5. Si el evento en el paso 3 es un evento de extinción, se elije un linaje vivo aleatorio en el árbol. Ese linaje ahora está muerto. Siempre que haya al menos un linaje vivo en el árbol, se regresa al paso 1; de lo contrario, se detiene la simulación.

Este procedimiento devuelve un árbol filogenético que incluye linajes vivos y muertos. Aunque hay formas mucho más eficientes de simular árboles (Stadler 2011).

Dos casos particulares del modelo nacimiento-muerte son los procesos Poisson ($\lambda_n = \lambda,\ \mu_n = \mu$) y el modelo de nacimiento puro ($\lambda_n = n\lambda,\ \mu_n = 0$).

3.2. Coalescente de Kingman

El coalescente es un modelo de distribución de la divergencia de genes en una genealogía. Es decir, es el proceso de rastrear las relaciones genealógicas de una muestra de cromosomas, o genes, hacia atrás en el tiempo a lo largo de las generaciones, en algún momento en el tiempo los linajes se unen o fusionan cuando se encuentran con sus ancestros comunes, este rastreo continua hasta que se alcanza el ancestro común más reciente (MRCA; Rosenberg y Nordborg 2002; Wakeley 2009). La teoría coalescente, originalmente formulada por Kingman (1982) como "$n - coalescente$" (para enfatizar la dependencia en el tamaño de la muestra, y que el resultado es para una población cuyo tamaño tiende a infinito, $N \to \infty$), se utiliza para estimar parámetros genéticos poblacionales, como el tamaño de la población, las tasas de migración y las tasas de recombinación en poblaciones. El modelo se deriva de un modelo genético poblacional simple propuesto por el genetista evolutivo Sewall Wright (1931).

Wright señaló que en una población finita de tamaño $2N$, la probabilidad de que dos copias de genes provengan de la misma copia en la generación anterior es $\frac{1}{2N}$ y en cada generación la probabilidad es la misma. El número de generaciones desde que dos genes compartieron por primera vez un ancestro común tiene una distribución $Geometrica(\frac{1}{2N})$ con una media $2N$. El resultado de Kingman (1982) es la extensión de este modelo en una población con k copias del gen. Retrocediendo en el tiempo, habrá varias generaciones hasta que dos o más de estas k copias tengan un ancestro común.

La primera copia provino de alguna copia de la generación anterior; la segunda tiene probabilidad $1 - \frac{1}{2N}$ de provenir de una copia diferente, de modo que ahora se representan dos copias de la generación anterior; la probabilidad de que la tercera copia provenga de una copia diferente de ambas es $1 - \frac{2}{2N}$ y la cuarta copia tiene probabilidad $1 - \frac{3}{2N}$ de provenir de una copia diferente de todas las anteriores. Entonces, definimos G_{kk} como la probabilidad de que k genes tengan k distintos antepasados en la generación anterior:

$$
\begin{aligned}
G_{kk} &= \Big(1-\frac{1}{2N}\Big)\Big(1-\frac{2}{2N}\Big)\Big(1-\frac{3}{2N}\Big)\cdots\Big(1-\frac{k-1}{2N}\Big) \\
&= 1-\Big(\frac{1+2+3+\cdots(k-1)}{2N}\Big)+\mathcal{O}\Big(\frac{1}{N^2}\Big) \\
&= 1-\Big(\frac{k(k-1)}{4N}\Big)+\mathcal{O}\Big(\frac{1}{N^2}\Big)
\end{aligned} \tag{7}
$$

Considerando que N (tamaño de la población) es más grande en relación con k (tamaño de la muestra), podemos ignorar $\mathcal{O}\Big(\frac{1}{N^2}\Big)$ y suponemos que el evento de que tres o más genes se fusionen (coalescan) en la misma generación es menos común en comparación con la fusión de dos genes. La probabilidad de observar un evento de coalescencia una generación atrás aumenta con el tamaño de la muestra.

Por lo tanto, la probabilidad de que al menos dos genes compartan un ancestro común en la generación anterior es

$$
1-G_{kk}=\Big(\frac{k(k-1)}{4N}\Big)+\mathcal{O}\Big(\frac{1}{N^2}\Big) \tag{8}
$$

Dado que esto es igual en cada generación, tenemos que el número de generaciones hasta que al menos dos genes en una muestra de k compartan un ancestro común tiene una distribución $Geometrica\Big(\frac{k(k-1)}{4N}\Big)$. Entonces, el tiempo medio desde el primer evento de coalescencia es

$$
\mathcal{E}(t_k)=\frac{4N}{k(k-1)} \tag{9}
$$

Este tiempo también puede ser aproximado con una distribución exponencial con la misma media. Entonces, los pasos para construir un árbol genealógico de k copias del gen son (Felsenstein 2004):

1. Dibujar una observación a partir de una distribución exponencial con media $\frac{4N}{k(k-1)}$. Este será el momento del primer evento coalescente (mirando desde el presente hacia atrás en el tiempo).

2. Se combinan dos linajes elegidos al azar.

3. k decrece en 1.

4. Si $k = 1$ se detiene el proceso, de otra manera se regresa al paso 1.

4. Modelo Beta-splitting y sus modificaciones

En esta sección introduciré dos modelos de generación de árboles filogenéticos aleatorios binarios basados en particiones de intervalos, asociados con la distribución Beta, los cuales definen el orden de las puntas (u hojas), sin considerar la longitud de las ramas. El primer modelo fue propuesto por Aldous (1996), una familia de cladogramas aleatorios de un parámetro (β), denominado modelo de división beta o beta-splitting.

Primero, para comprender mejor los siguientes modelos e introducir el efecto del parámetro β y la influencia que tiene sobre las particiones, la forma y equilibrio de los árboles generados, daré una definición de la distribución Beta.

La distribución Beta es una distribución de probabilidad continua que está limitada por el intervalo $[0, 1]$, sus parámetros a y b, aparecen como exponentes de la variable aleatoria y controlan la forma de la distribución. Entonces, decimos que la variable aleatoria continua R tiene una distribución Beta con parámetros $a > 0$ y $b > 0$, es decir, $R \sim Beta(a, b)$, cuando su función de densidad es

$$f_R(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \quad (x \in [0,1])$$

donde $\frac{1}{B(a,b)}$ es una constante normalizadora para hacer que la función integre a 1. $B(a, b)$ es conocida como la función beta dada por la siguente integral

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

y esta relacionada con la función gamma, a través de la identidad

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Usaremos esto para ver algunos ejemplos de la distribución Beta y como cambia su forma de acuerdo a los valores que toman los parámetros.

En las Figuras 5 y 6 podemos notar que la distribución se centra aproximadamente en $\frac{a}{(a+b)}$, la media de la distribución. Además, cuanto mayores sean los valores de a y b, menor será la varianza de la distribución con respecto a la media, $Var(R) = \frac{ab}{(a+b+1)(a+b)^2}$. También podemos ver que para valores grandes de los parámetros, la distribución toma una forma más picuda. Para valores de a mayores a b vemos que la curva de la distribución se sesga a la derecha, y cuando a es menor a b se sesga a la izquierda. En la Figura 5, observamos que la

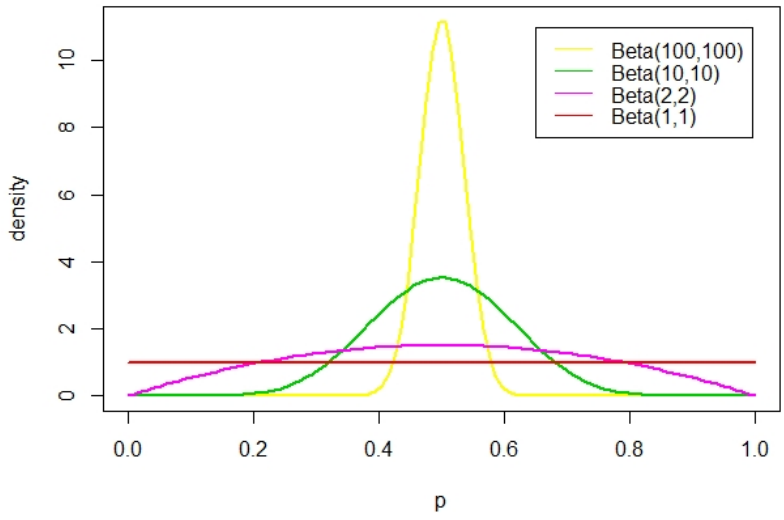

Figura 5: $a = b$, ambos ≥ 1.

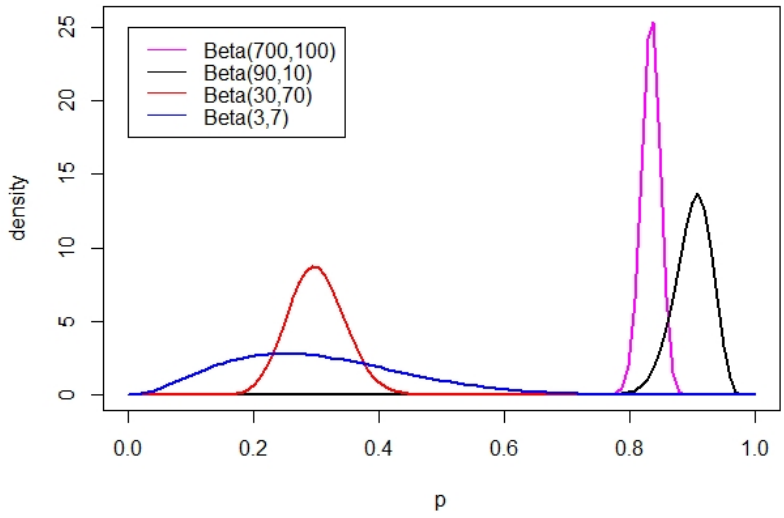

Figura 6: $a \neq b$, ambos ≥ 1.

distribución Beta se reduce a la distribución uniforme, $Unif(0, 1)$, cuando $a = b = 1$.

Los parámetros a y b también pueden ser menores que 1, pero la distribución tiene una forma diferente, como se puede ver en la Figura 7. Específicamente, si $a < 1$, entonces hay un pico en 0, y si $b < 1$, entonces hay un pico en 1 y si ambos son menores a 1, entonces la distribución tiene forma de U.

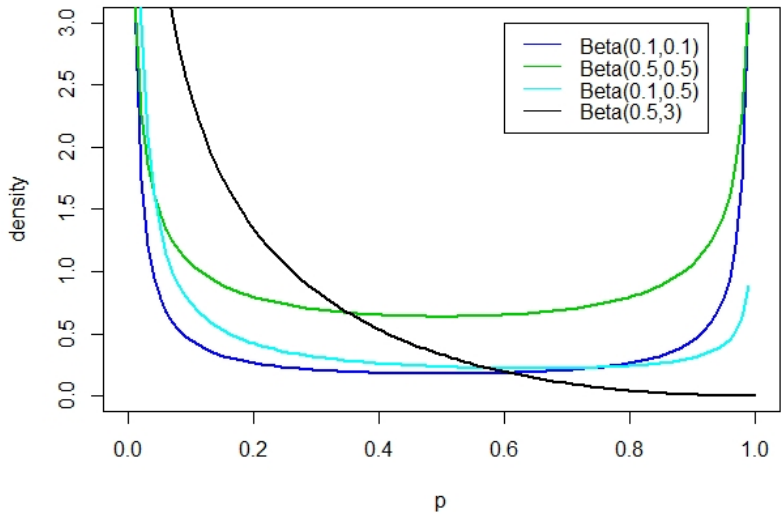

Figura 7: a, b, ≤ 1.

4.1. Modelo β-splitting

El modelo β-splitting, propuesto por Aldous (1996), es una construcción matemática bastante general de un árbol aleatorio con un número específico de puntas, n especies etiquetadas $\{1, 2, \ldots, n\}$ en un intervalo unitario $[0, 1]$ en posiciones aleatorias uniformes. Entonces, el modelo se define como sigue, se divide el intervalo en un punto aleatorio elegido con alguna densidad de probabilidad f que satisface la condición de simetría $f(x) = f(1 - x)$ para $x \in (0, 1)$.

En el modelo β-splitting f es la densidad de una distribución $Beta(\beta + 1, \beta + 1)$. La primera partición del intervalo $[0, 1]$ se determina dibujando un punto aleatorio R, definido por el parámetro $\beta \in [-1, \infty)$, con la distribución Beta simétrica, $R \sim Beta(\beta + 1, \beta + 1)$. Luego, se siguen dividiendo los subintervalos recursivamente. Un subintervalo X, de ancho $|X|$ se divide en dos subintervalos separados, X_{left} y X_{right}, con longitudes $|X_{left}| = R|X|$ y $|X_{right}| = (1 - R)|X|$, en distribución. Entonces, el algoritmo, ilustrado en la Figura 8, comienza con n variables aleatorias uniformes e independientes $(U_i)_{i\in\{1,\ldots,n\}}$ en el intervalo $[0, 1]$, representando las n hojas. El intervalo $[0, 1]$ se subdivide secuencialmente, hasta que todas las marcas estén en particiones distintas.

Durante la primera partición, si tenemos n variables aleatorias independientes $\sim Unif(0, 1)$, el número Y de ellas que caen del lado izquierdo del split (definido por el parámetro β) es (condicionalmente a $R = x$) $Binomial(n, x)$, entonces $Y|R = x \sim Bin(n, x)$.

En general se tiene

$$\begin{aligned}\mathbb{P}(Y=i) &= \int \mathbb{P}(Y=i|R=x)d\mathbb{P}_R(x) \\ &= \int \binom{n}{i} x^i(1-x)^{(n-i)}d\mathbb{P}_R(x) \\ &= \binom{n}{i}\int x^i(1-x)^{(n-i)}\frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}x^\beta(1-x)^\beta dx \\ &= \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}\binom{n}{i}\int x^i(1-x)^{(n-i)}x^\beta(1-x)^\beta dx.\end{aligned} \tag{10}$$

Como se prohibe que los n puntos caigan del lado izquierdo (o del lado derecho), utilizaremos la constante de renormalización

$$\begin{aligned}a_n &= \sum_{i=1}^{n-1}\binom{n}{i}\int_0^1 x^i(1-x)^{n-i}x^\beta(1-x)^\beta dx \\ &= \int_0^1 \Big(\sum_{i=1}^{n-1}\binom{n}{i}x^i(1-x)^{n-i}\Big)x^\beta(1-x)^\beta dx \\ &= \int_0^1 \big(1-x^n-(1-x)^n\big)x^\beta(1-x)^\beta dx.\end{aligned} \tag{11}$$

Como $1-x^n-(1-x)^n \sim nx$ cerca de 0, vemos que este modelo se puede extender a $\beta > -2$.

En este modelo los nodos internos no están etiquetados, por lo que no se registra el orden de las particiones, se dice que este árbol es no-clasificado. El parámetro $-2 < \beta < \infty$ modulará la forma y el balance del árbol, como vimos en las Figuras 5, 6 y 7, dependiendo de los valores de los parámetros se definirá el centro y el sesgo de la curva de la distribución y esto es también lo que define dónde será la siguiente partición en el intervalo, entre más alto el valor del parámetro β más balanceado será el árbol generado; si $\beta \to -2$, el modelo genera el árbol desbalanceado perfecto con probabilidad de uno, y con $\beta \to \infty$, el modelo genera el árbol balanceado perfecto con probabilidad de uno (Kim et al. 2019).

4.2. Un nuevo modelo de partición

El siguiente modelo es desarrollado en el artículo *Ranked Tree Shapes, Nonrandom Extinctions, and the Loss of Phylogenetic Diversity* por Maliet et al. (2018). Los autores investigan cómo la pérdida de diversidad filogenética está influenciada por dos factores: (i) la forma del árbol (ranked shape), caracterizada por la relación entre la riqueza de clados y su edad y (ii)

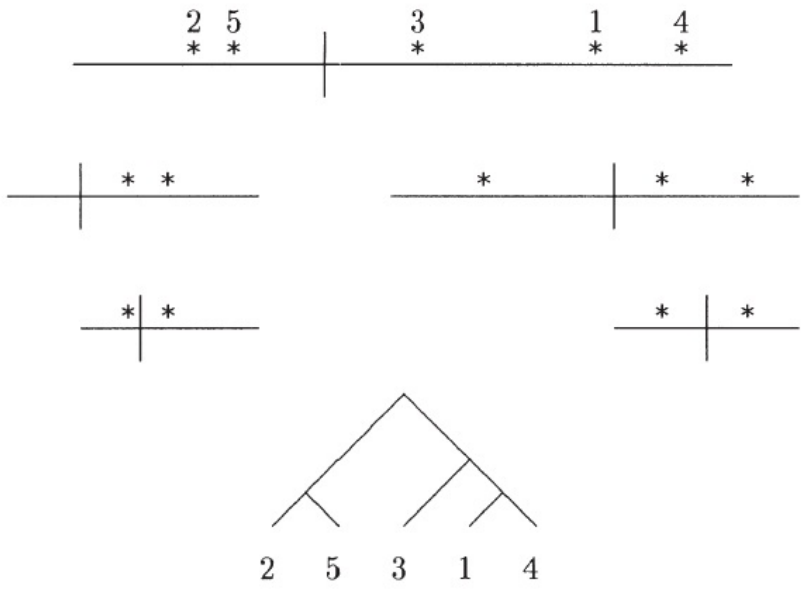

Figura 8: Construccion de la division de intervalos en el modelo beta-splitting. Obtenido de Aldous (1996).

las extinciones no aleatorias, caracterizadas por la relación entre la riqueza de clados y los riesgos de extinción dentro de ellos.

La propuesta de los autores es un modelo de tres parámetros que genera formas de árboles clasificados aleatorios, es decir las particiones tienen un orden de generación. Este modelo puede considerarse una extensión del modelo beta-splitting de Aldous (1996, 2001) con dos parámetros adicionales: un parámetro $\alpha \in (-\infty, \infty)$, que cuantifica la relación entre la riqueza de un clado y su edad relativa (es decir, el rango de aparición de su nodo raíz), denominado índice de edad-riqueza, y otro parámetro $\eta \in [0, \infty)$ que cuantifica la relación entre la riqueza de un clado y su abundancia relativa (la suma de las abundancias de las especies consideradas dividida por la suma de las abundancias de todas las especies existentes en la filogenia). Exploran la tasa de disminución de la diversidad filogenética a medida que las especies se extinguen secuencialmente, basado en datos simulados bajo la variación de los tres parámetros en un rango de sus posibles valores.

El algoritmo, ilustrado en la Figura 9, comienza con n variables aleatorias uniformes e independientes $(U_i)_{i\in\{1,\ldots,n\}}$ en el intervalo $[0, 1]$. El intervalo $[0, 1]$ se subdivide secuencialmente, hasta que todas las marcas estén en particiones distintas.

La primera partición del intervalo $[0, 1]$ se determina dibujando un punto aleatorio R, definido por el parámetro $\beta \in (-2, \infty)$, con la misma distribución Beta simétrica, $R \sim Beta(\beta + 1, \beta + 1)$, y función de densidad que el modelo β-splitting de Aldous (1996, 2001). Luego, se siguen dividiendo los subintervalos recursivamente. Un subintervalo X, de ancho $|X|$ se divide en dos subintervalos separados, X_{left} y X_{right}, con longitudes $|X_{left}| = R|X|$ y

$|X_{right}| = (1 - R)|X|$, en distribución.

A cada intervalo X de la partición, con al menos dos marcas, se le da un peso de $|X|^{\alpha}$. Uno de estos intervalos se selecciona con una probabilidad proporcional a su peso, es decir, si $\alpha > 0$ entonces el intervalo de mayor longitud tiene más probabilidad de ser seleccionado en la siguiente partición, y cuando $\alpha < 0$ sucede lo contrario (el efecto del parámetro se explicará más adelante).

El algoritmo continúa generando valores con distribución beta para bi-particionar las hojas (intervalos) y elige el orden proporcional a su número de descendientes hasta que el intervalo $[0, 1]$ se divide en n intervalos, cada uno correspondiente a una U_i. Lo anterior se realiza incluso si un subintervalo no contiene ninguna marca entre $(U_i)_{i\in\{1,\dots,n\}}$, esto corresponde a un subárbol sin especies muestreadas.

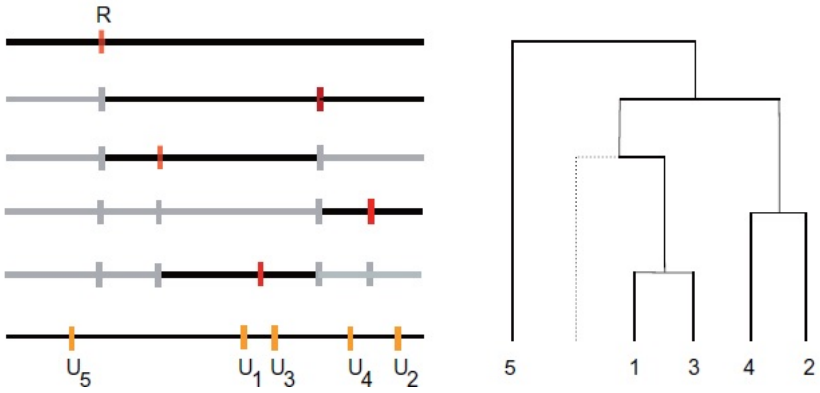

Figura 9: Ilustración de la generación de árboles ranqueados con el modelo alpha-beta splitting. (1) Cinco marcas aleatorias $(U_i)_{i\in 1,\dots,5}$ se dibujan uniformemente en el intervalo $[0, 1]$ (las marcas en la línea inferior). (2) Conforme avanza el tiempo (tiempo fluye hacia abajo), seleccionamos aleatoriamente un intervalo X (en negro). Luego, dibujamos una variable aleatoria R en una distribución Beta con parámetros $(\beta+1, \beta+1)$ (marca roja), y dividimos el intervalo seleccionado en dos subintervalos. (3) Se repite este proceso a lo largo del tiempo hasta que todos los intervalos contengan solo una marca. Obtenido de Maliet et al. (2018).

Como se explicó arriba, en este modelo la forma y clasificación (orden de los nodos) del árbol lo definen los parámetros β y α, respectivamente, similar al modelo β-splitting. De esta manera, si el valor de β es pequeño (cercano a -2) la partición de los intervalos será cerca de las extremidades, como resultado será un árbol desbalanceado; si el valor es grande la partición se acercará a la mitad y el árbol será balanceado.

En el caso de $\alpha \in (-\infty, \infty)$, aunque sea modificado, se mantiene la forma del árbol. Cuando α tiene valores pequeños, particularmente $\alpha < 0$, una vez que se hizo la partición del intervalo por β, la partición de menor longitud tiene más probabilidad de ser seleccionada, y el subárbol con el menor número de especies estará cerca del nodo raíz, es decir, las especies son más antiguas. Por el contrario, cuando los valores de α son grandes, las particiones de mayor

longitud tienen más probabilidad de ser seleccionados, y el subárbol con menor número de especies estará cerca de las puntas, es decir, las especies serán más recientes. El efecto de los parámetros α y β se observa claramente en la Figura 10, en el árbol de la esquina superior izquierda tenemos un valor grande de β y pequeño de α, vemos que el árbol es balanceado y el subárbol derecho tiene menor número de especies, las cuales tienen mayor edad, mientras que el subárbol izquierdo tiene el mayor número de especies que son de menor edad. Por otro lado, en el árbol de la esquina inferior derecha tenemos valores pequeños de β y valores grandes de α, el árbol es claramente desbalanceado, en el subárbol izquierdo se encuentran la mayoría de las especies, que aparecieron más atrás en el tiempo, mientras que en el derecho solo están dos especies, que aparecieron recientemente en el tiempo.

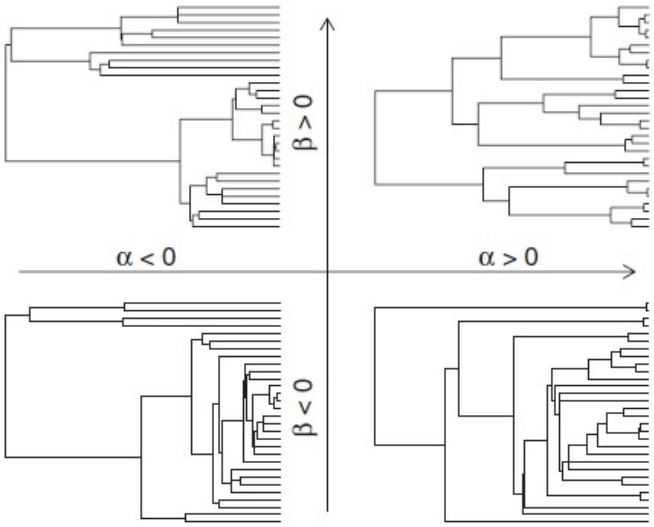

Figura 10: Árboles filogenéticos mostrando el efecto de diferentes valores del parámetro $\beta \in (-2, \infty)$ que determina el equilibrio del árbol, y $\alpha \in (-\infty, +\infty)$ que establece la relación entre la riqueza de especies de un clado y su edad relativa. Obtenido de Maliet et al. (2018).

En el caso de que $\alpha = 1$ se convierte en el modelo beta-splitting, pues se anula el efecto de α.

En cuanto al parámetro $\eta \geq 0$, también llamado índice de abundancia-riqueza, permitirá determinar el orden de extinción de las especies en función de su abundancia, es decir, las especies menos abundantes se extinguen primero, mientras que las especies más abundantes se extinguen al final. El parámetro se incorpora al modelo de la siguiente manera:

Cuando un intervalo es particionado en dos intervalos, $|X_{left}| = R|X|$ y $|X_{right}| = (1-R)|X|$, significa la formación de dos subárboles, a los cuales se les asigna una parte de la abundancia relativa A_X del clado como sigue,

$$A_{X_{left}} = \frac{|X_{left}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A_X = \frac{R^\eta}{R^\eta + (1-R)^\eta} A_X$$
$$A_{X_{right}} = \frac{|X_{right}|^\eta}{|X_{left}|^\eta + |X_{right}|^\eta} A_X = \frac{(1-R)^\eta}{R^\eta + (1-R)^\eta} A_X. \tag{12}$$

Entonces, cuando $\eta < 1$ las especies en clados más grandes tienen abundancias más pequeñas, en promedio, y por lo tanto mayores riesgos de extinción. Sin embargo, cuando $\eta = 1$ todas las especies tienen, en promedio, la misma abundancia $(\frac{1}{n})$, todas las especies tienen la misma probabilidad de extinguirse. En el caso de $\eta > 1$ las especies en clados más grandes tienen abundancias más grandes, y menor riesgo de extinción, tal como se presenta en la Figura 11, donde el grosor de los círculos es proporcional a la abundancia de las especies. En la imagen, en el primer árbol, donde $\eta = 0{,}2$, los círculos son más grandes en ambos extremos, que son subárboles con el menor número de especies, a diferencia del árbol donde $\eta = 3$, el subárbol izquierdo es el más grande y tiene especies con abundancias más grandes.

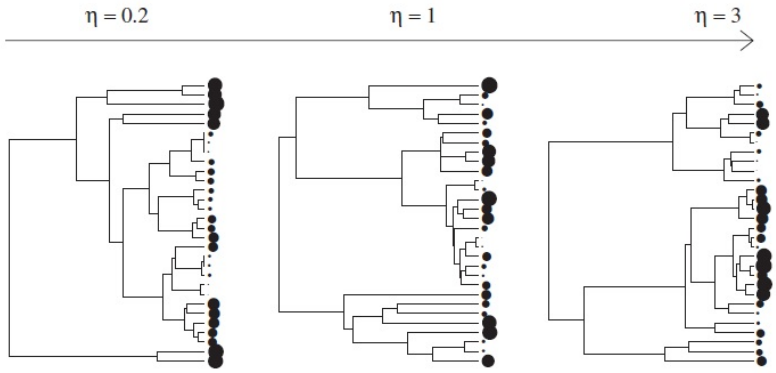

Figura 11: Árboles filogenéticos que muestran la distribución de frecuencias de especies a lo largo de las puntas de para diferentes valores del índice de abundancia-riqueza, η. Los tamaños de puntos clasifican las especies según su frecuencia, puntos más grandes indican especies más abundantes.Obtenido de Maliet et al. (2018).

Los autores proponen un algoritmo de inferencia de Monte-Carlo que permite la estimación de máxima verosimilitud de los parámetros β, α y η.

4.3. Inferencia de parámetros

La inferencia de los parámetros propuesta es por máxima verosimilitud. En el caso de β esta versión del modelo solo permite la simulación de árboles con $\beta > -1$, ya que la distribución

beta solo se define para valores positivos $(\beta + 1, \beta + 1)$. Dado que la probabilidad de la forma del árbol bajo este modelo es la misma que en el modelo de Aldous (independiente de α y η), podemos usarlo para estimar β (Aldous 1996). Esta inferencia proviene de que la verosimilitud de β, $L(i, n|\beta)$ es

$$a_n^{-1} \binom{n}{i} \int x^{i+\beta}(1-x)^{n-i+\beta} dx \tag{13}$$

donde la constante de normalización, a_n, se define de la misma manera que en (11).

En cuanto a la verosimilitud de α, solamente cuando la longitud de los intervalos es conocida, tenemos que para cada $k \in N$, donde k es el $k$ésimo paso en que se escoge entre n_k intervalos con respectivas longitudes $|X_1^k|, \ldots, |X_k^k|$ la probabilidad de seleccionar X_i^k está dada por

$$p_k(X_i^k|\alpha) = \frac{|X_i^k|^\alpha}{\sum_{i=1}^{n_k} |X_j^k|^\alpha} \quad para \quad 1 \leq i \leq n_k. \tag{14}$$

Si X^k denota el $k$ésimo intervalo seleccionado entonces $X^k \sim p_k(\cdot|\alpha)$ y las variables $(X^k)_{k \in N}$ son independientes. De esta forma, si se tienen datos $X^1 = X_{i_1}^1, \ldots, X_l = X_{i_l}^l$ la verosimilitud para α bajo el modelo β-splitting es

$$L(\alpha|\bar{X}) = \prod_{k=1}^{l} p_k(X_{i_k}^k|\alpha). \tag{15}$$

Aún cuando no se puede obtener una fórmula explícita para el estimador de máxima verosimilitud, éste se puede aproximar mediante métodos numéricos.

Debido a la dificultad para hacer inferencias explícitas sobre α y η (la inferencia del parámetro β es directa y más sencilla) los autores utilizaron el procedimiento de aumento de datos de Monte-Carlo.

Los métodos de Monte-Carlo son métodos numéricos basados en simulación estocástica que son útiles para hacer estimaciones en modelos probabilísticos. En este caso se utilizará para hacer estimación sobre parámetros del modelo a través de la simulación de una gran cantidad de árboles filogenéticos aleatorios. Suponiendo que tenemos una muestra de árboles filogenéticos a los cuales les queremos ajustar un modelo alpha-beta splitting, con parámetros α, β y η, la idea para llevar a cabo la estimación sobre los parámetros primero se comienza por dar un estimador puntual de β, para el cual tenemos una verosimilitud explícita, y una vez que se cuenta con este parámetro, para determinar α y η.

Además, el aumento de datos se refiere a la introducción de datos no observados por medio de métodos de simulación, en este caso, los datos no observados son las particiones de inter-

valos sin muestrear, es decir, las especies no muestreadas. La idea es la construcción del árbol bajo el modelo alpha-beta splitting suponiendo que conocemos donde tenemos especies sin muestrear. Entonces, con la siguiente densidad condicional podemos simular la construcción del árbol basándonos únicamente en su ramificación (longitud de intervalos),

$$P(R \in dx|Y = k) = \frac{x^{\beta+k}(1-x)^{\beta+n-k}}{\int_0^1 y^{\beta+k}(1-y)^{\beta+n-k}dy}dx. \tag{16}$$

Se trata de la ley condicional del punto de partida del intervalo, dado que hay k hojas del lado izquierdo, qué tan probable es que x esté en el intervalo (0,1). Entonces, podemos reconstruir una simulación de la longitud de los intervalos (una vez conocida la longitud del intervalo la verosimilitud de α, presentada anteriormente, es explícita).

Maliet et al. (2018) simularon 20 árboles con 20, 50 y 100 puntas (especies) para todas las posibles combinaciones de $\alpha \in (-1, 0, 1, 2)$, $\beta \in (-1, 0, 1)$ y $\eta \in (0{,}2, 0{,}5, 1, 1{,}5, 2)$, obteniendo un total de 3600 árboles. Luego infirieron los parámetros del modelo en estos árboles y los compararon con los valores usados en las simulaciones.

A continuación, presento una serie de árboles simulados bajo la propuesta de los autores, así como la comparación entre los valores de los parámetros usados en las simulaciones y los resultados de las inferencias. Las simulaciones de los árboles se presentan para 20, 50 y 100 puntas.

En cuanto al balance de los árboles, podemos ver en las Figuras 12 y 13, donde el valor de β es 1, que se trata de árboles balanceados, la partición de los intervalos es cerca de la mitad sin importar el número de puntas. Mientras que en las Figuras 14, 15 y 16 en las que los valores de β son 0 y -1 la partición de los intervalos es cerca de los extremos, se trata de árboles desbalanceados. El desbalance se puede apreciar mucho más en la Figura 15 con $\beta = -1$, sin importar el número de puntas en la simulación, uno de los extremos contiene solo una o dos puntas. En el caso del efecto de α, la Figura 12 se simuló con $\alpha = 1$, como se mencionó anteriormente, este es el caso en el que se anula el efecto del parámetro y el modelo se convierte en el modelo beta-splitting propuesto por Aldous (1996). Sin embargo, en las Figuras 13 y 14 donde los valores de α son 0 y 2, respectivamente, vemos que las particiones de mayor longitud, es decir, las que cuentan con mayor número de puntas/especies están más cerca de las puntas (son más jóvenes), esto se aprecia mejor en las simulaciones con 100 puntas. Por el contrario, las Figuras 15 y 16, donde $\alpha = -1$, vemos que los subárboles con menor número de puntas son los que están más cerca de la raíz (son más antiguas). Sobre el efecto de η, en las Figuras 12 y 13 donde $\eta = 1$, todas las puntas tienen la misma abundancia, podemos ver que el tamaño de los cuadros está distribuído de manera uniforme, es decir, todas las especies tienen la misma probabilidad de extinguirse. Pero en las Figuras 14 y 16, donde η toma valores de 2 y 1,5, vemos que los cuadros más grandes se concentran en el subárbol con mayor número de especies.

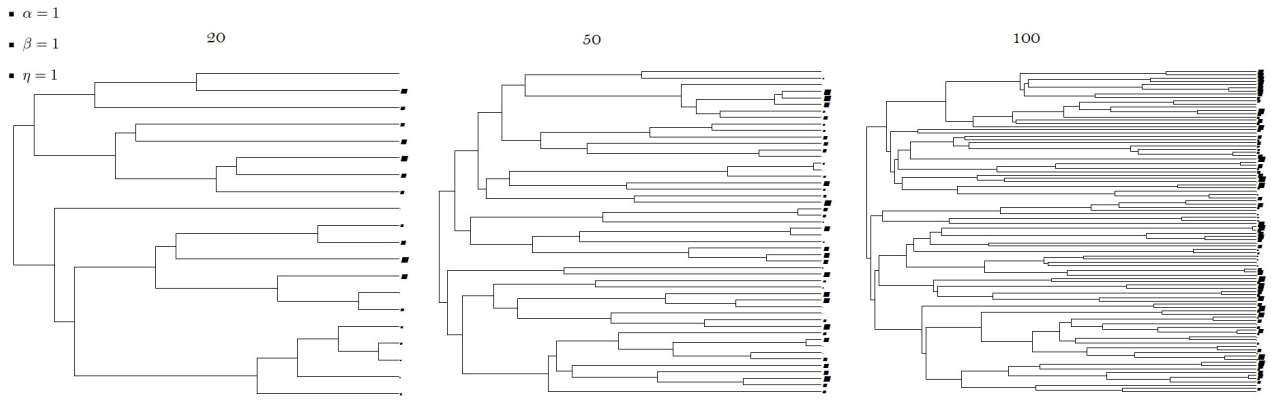

Figura 12: Árboles filogenéticos simulados para diferentes valores de número de especies N. Valores de los parámetros: $\beta = 1$, $\alpha = 1$ $\eta = 1$, y el parámetro de aproximación $\epsilon = 0{,}001$.

N	$\alpha = 1$	$\beta = 1$	$\eta = 1$
20	0.45	1.56	2.03
50	1.90	2.10	1.53
100	1.15	0.95	1.04

Cuadro 1: Inferencias de los parámetros de la Figura 12.

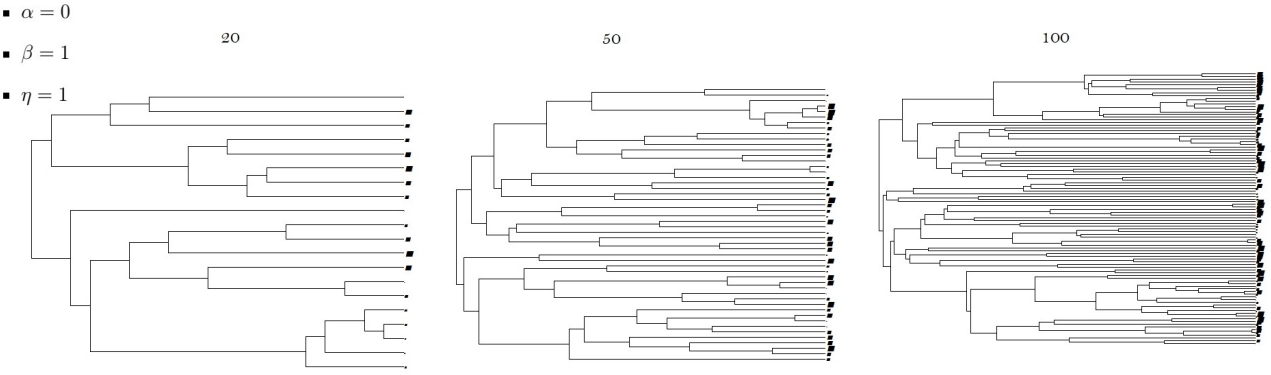

Figura 13: Árboles filogenéticos simulados para diferentes valores de número de especies N. Valores de los parámetros: $\beta = 1$, $\alpha = 0$ $\eta = 1$, y el parámetro de aproximación $\epsilon = 0{,}001$.

N	$\alpha = 0$	$\beta = 1$	$\eta = 1$
20	0.04	1.56	1.14
50	1.03	2.11	3.49
100	0.07	0.95	0.87

Cuadro 2: Inferencias de los parámetros de la Figura 13.

Podemos observar los resultados de las inferencias realizadas con el algoritmo de inferencia de Monte-Carlo por aumento de datos en los Cuadros 1-5 y vemos que funciona razonablemente bien en filogenias con más de 50 puntas. En el Cuadro 2 vemos que para $\beta = 0$ y

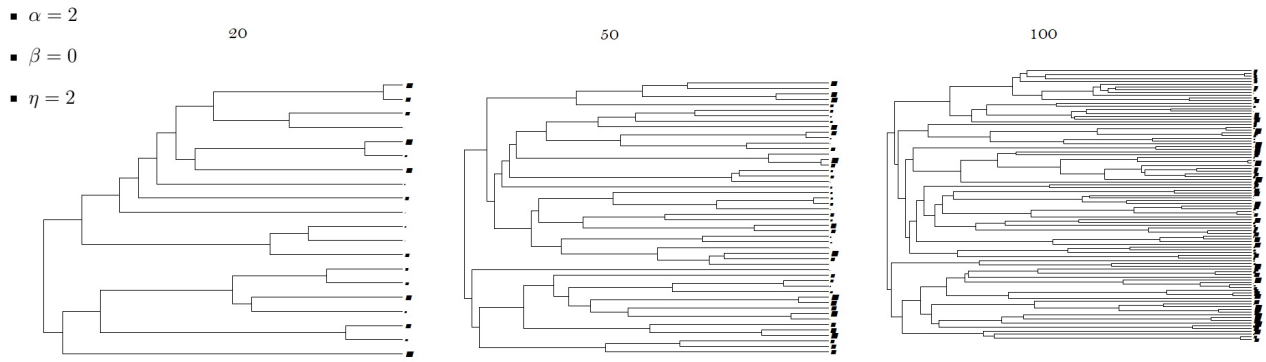

Figura 14: Árboles filogenéticos simulados para diferentes valores de número de especies N. Valores de los parámetros: $\beta = 0$, $\alpha = 2$ $\eta = 2$, y el parámetro de aproximación $\epsilon = 0{,}001$.

N	$\alpha = 2$	$\beta = 0$	$\eta = 2$
20	1.20	-0.75	1.70
50	2.14	0.07	2.08
100	2.06	0.24	2.14

Cuadro 3: Inferencias de los parámetros de la Figura 14.

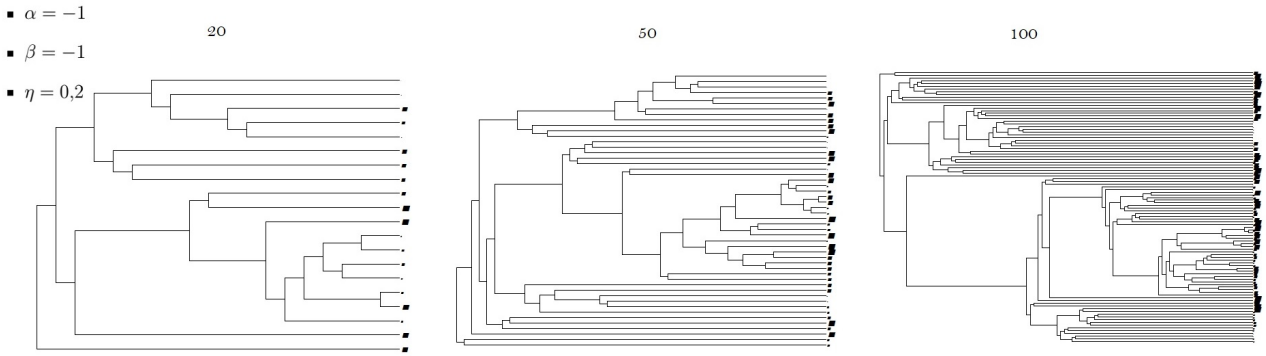

Figura 15: Árboles filogenéticos simulados para diferentes valores de número de especies N. Valores de los parámetros: $\beta = -1$ $\alpha = -1$ $\eta = 0{,}2$, y el parámetro de aproximación $\epsilon = 0{,}001$.

N	$\alpha = -1$	$\beta = -1$	$\eta = 0{,}2$
20	-0.42	-1.25	9.57
50	-0.77	0.07	2.08
100	-0.73	-0.94	9.76

Cuadro 4: Inferencias de los parámetros de la Figura 15.

$\alpha \leq 0$, las inferencias de los tres parámetros son buenas en el caso de las simulaciones con 100 puntas, pero son sobreestimados en el caso de 20 y 50 puntas. En el caso del Cuadro 3, para $\beta = 0$ y $\alpha \geq 0$, η se subestima cuando el numero de puntas es 20, pero las inferencias

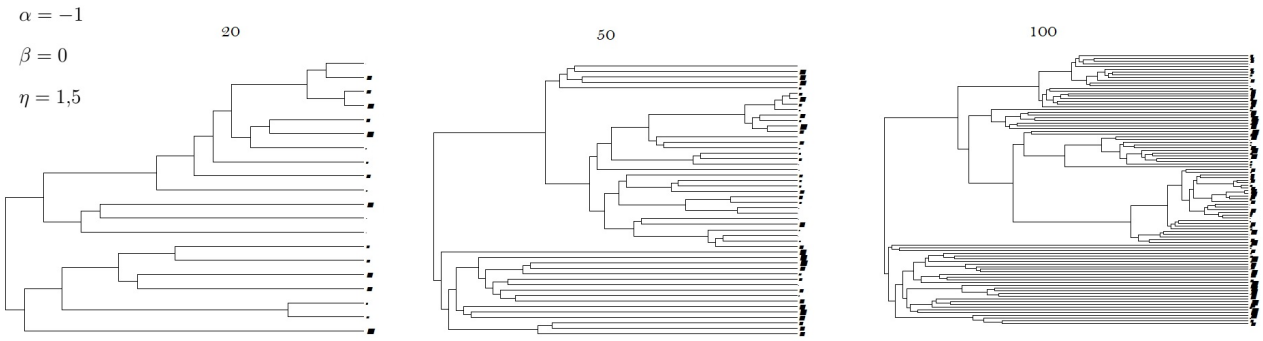

Figura 16: Árboles filogenéticos simulados para diferentes valores de número de especies N. Valores de los parámetros: $\beta = 0$ $\alpha = -1$ $\eta = 1{,}5$, y el parámetro de aproximación $\epsilon = 0{,}001$.

N	$\alpha = -1$	$\beta = 0$	$\eta = 1{,}5$
20	-0.73	-0.75	1.13
50	-1.42	0.08	1.65
100	-0.94	0.24	1.46

Cuadro 5: Inferencias de los parámetros de la Figura 16.

son buenas, a partir de 50 puntas. Algo inesperado ocurrió cuando $\eta = 0{,}2$ (Cuadro 4), en los tres casos η es inusualmente sobreestimado, mientras que las inferencias de los otros dos parámetros son relativamente buenas a partir de 50 puntas. Sin embargo, en cualquier otro caso donde $\eta > 0{,}2$ el algoritmo de inferencia devuelve buenas estimaciones generales de α y η. En cuanto a β, su estimación en árboles con al menos 50 puntas es precisa, ya que la inferencia del parámetro es directa.

En conclusión, Maliet et al. (2018) presentan un nuevo modelo para formas de árbol clasificadas aleatoriamente con un número arbitrario de puntas. Este modelo presenta dos parámetros, β y α que ajustan respectivamente el equilibrio del árbol y las edades relativas de sus nodos. De manera general, en las Figuras 12-16 podemos observar que los árboles con $\beta \leq 0$ están desequilibrados y los árboles con $\beta > 0$ están equilibrados. Además, cualquiera que sea el valor de α, la forma del árbol es la misma que en el modelo de Aldous (Aldous 1996, 2001). Los clados con mayor número de especies tienden a coalescer en lo profundo del árbol cuando $\alpha > 0$ y son menos profundos que los clados con menor número especies cuando $\alpha < 0$. Por otro lado, este modelo satisface la propiedad de consistencia de muestreo, que se refiere a que un árbol con n puntas tiene la misma distribución que un árbol con $n + 1$ puntas con una punta eliminada al azar. Esta propiedad es esencial para asegurar la robustez del modelo cuando nos enfrentamos a un muestreo incompleto de taxones. La incorporación del índice de abundancia-riqueza η dentro del marco proporcionado por el modelo de Aldous permite simular un rasgo que covaría con la forma de la filogenia (definida por los parámetros β y α). Y cuando $\eta > 1$, las especies más abundantes se encuentran en los clados con más especies, mientras que cuando $\eta < 1$ las especies más abundantes se encuentran en los clados con menos especies y cuando $\eta = 1$ todas las especies tienen la misma abundancia en promedio.

Esto nos hace suponer que las extinciones ocurren secuencialmente en orden de abundancia creciente. En este caso, consideramos que el riesgo de extinción se debe a la rareza de una especie (que tan abundante es una especie).

5. Bibliografía

Referencias

[1] Aldous D. (1996). *Probability distributions on cladograms.* In In Aldous D, Pemantle R, editors, *Random Discrete Structures*. pages 1–18, New York, NY.

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

[3] Blum M. G. B. & Francois O. (2006). *Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance.* Systematic Biology, 55, 685–691.

[4] Faith D. P. (1992). *Conservation evaluation and phylogenetic diversity*. Biological Conservation, 61, 1-10.

[5] Faller B., Pardi F. & Steel M. (2008) *Distribution of phylogenetic diversity under random extinction*. Journal of Theoretical Biology.

[6] Felsenstein J. (2004). *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts.

[7] Ford D. J.(2005). *Probabilities on cladograms: introduction to the alpha model*. En preparación.

[8] Heard S. B. (1992). *Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees*. Evolution, 46, 1818-1826.

[9] Hollering B. & Sullivant S. (2019). *Exchangeable and Sampling Consistent Distributions on Rooted Binary Trees.* arXiv preprint.

[10] Kim J., Rosenberg N. A. & Palacios J. A. (2019). *A Metric Space of Ranked Tree Shapes and Ranked Genealogies*. bioRxiv.

[11] Kingman J. (1982). *The coalescent*. Stochastic Processes and their Applications, 13: 235–248.

[12] Maliet O., Gascuel F. & Lambert A. (2018). *Ranked Tree Shapes, Nonrandom Extinctions, and the Loss of Phylogenetic Diversity*. Systematic Biology, 67(6), 1025–1040.

[13] Nee S. & May R. M. (1997). *Extinction and the Loss of Evolutionary History*. Science, 278, 692–694.

[14] Redding D. W. & Mooers A. O. (2006). *Incorporating Evolutionary Measures into Conservation Prioritization*. Conservation Biology, 20(6), 1670–1678.

[15] Rosenberg N. A. & Nordborg M. (2002). *Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.* Nature Reviews Genetics, 3, 380–390.

[16] Sainudiin R. & Véber A. (2016). *A beta-splitting model for evolutionary trees.* Royal Society Open Science, 3, 160016.

[17] Stadler T. (2011). *Simulating trees with a fixed number of extant species.* Systematic Biology, 60, 676–684.

[18] Veron S., Davies J. T., Cadotte M. W., Clergeau P. & Pavoine S. (2015) *Predicting loss of evolutionary history: Where are we?* Biological Reviews.

[19] Wakeley J. (2009). *Coalescent Theory: An Introduction.* Roberts & Co. Publishers, Greenwood Village, Colorado.

[20] Wright S. (1931). *Evolution in Mendelian populations.* Genetics, 16, 97–159.