Tequio *4*(11), 2021: 81-89
ISSN: 2594-0546

Métodos de reconstrucción filogenética II: inferencia bayesiana

Methods for phylogenetic reconstruction II: Bayesian inference

Pablo Duchen[1*]

Fecha de recepción: 6 de noviembre de 2020
Fecha de aceptación: 22 de enero de 2021

Resumen - La inferencia bayesiana como modelo de reconstrucción filogenética es muy utilizada en la actualidad. La ventaja de este método es la generación directa de probabilidades posteriores para cada clado en la filogenia final, por lo cual no se requiere de bootstrapping como medida de incertidumbre. Además, la inferencia bayesiana se presta perfectamente para la datación de filogenias por medio de relojes moleculares. En este trabajo se describen los principios de este método, comenzando por el teorema de Bayes; posteriormente se caracteriza el uso del algoritmo de Metropolis-Hastings para el muestreo de las topologías más probables y se le ilustra con un ejemplo sencillo. Se finaliza mencionando los programas más usados actualmente.

Palabras clave: Teorema de Bayes, Metropolis-Hastings, MCMC.

Abstract - Phylogenetic reconstruction through Bayesian inference is currently widely used. The main advantage of this method is the direct output of posterior probabilities for each clade on the final phylogeny. Thus, it does not require bootstrapping as a measure of uncertainty. Moreover, Bayesian inference is perfectly fit for dating phylogenies through molecular clocks. In this paper, the basics of Bayesian inference applied to phylogenetic reconstruction are described, starting with an explanation of Bayes' theorem. Then, the use of the Metropolis-Hastings algorithm to sample topologies from the posterior distribution is characterized and illustrated through a simple example. At the end, there is a mention of the software used for Bayesian phylogeny reconstruction.

Keywords: Bayes' theorem, Metropolis-Hastings, MCMC.

Introducción

Se presenta ahora el método bayesiano para reconstrucción filogenética. Como se verá a continuación, algunos elementos (como el cálculo de verosimilitudes y los modelos de mutación de ADN) también se utilizarán aquí; para no repetir su descripción, el lector deberá referirse a la primera parte de esta revisión.

En la inferencia bayesiana para la reconstrucción filogenética, el objetivo es encontrar el árbol con la mayor probabilidad posterior. Dicho cálculo va a depender de asumir una probabilidad a priori para cada árbol

[1] Departamento de Biología Computacional, Universidad de Lausana, Suiza.
Correos electrónicos: pablo.duchenbocangel@unil.ch, pduchen@gmail.com. ORCID: 0000-0002-9318-5002

y del uso de métodos markovianos (cadenas markovianas). De manera general, las cadenas markovianas son procesos estocásticos que describen una secuencia de eventos donde la probabilidad de un evento actual depende únicamente del anterior. En la inferencia bayesiana las cadenas markovianas se emplean para explorar el espacio de filogenias posibles.

La inferencia filogenética por medio de métodos bayesianos fue introducida por Rannala & Yang (1996), mientras que las extensiones markovianas fueron agregadas independientemente por Yang & Rannala (1997); Mau & Newton (1997) y Li, Pearl & Doss (2000). La base fundamental de toda inferencia bayesiana radica en el teorema de Bayes, el cual describimos a continuación.

Teorema de Bayes

Se debe iniciar por definir nuestros datos y parámetros a estimar. Dado un alineamiento *D* de secuencias de ADN para un número *n* de especies, el objetivo es encontrar el árbol *T* que mejor describa a dicho alineamiento. En esta revisión se usarán los términos filogenia y árbol indistintamente para referirse a *T*, al igual que los términos alineamiento o datos para aludir a *D*.

El teorema de Bayes en inferencia filogenética se presta fácilmente para calcular *T* dado un alineamiento *D*. Bajo este teorema, la probabilidad posterior de *T* es:

$$P(T|D) = \frac{P(T)P(D|T)}{P(D)}, \quad (1)$$

donde *P(T)* es la probabilidad *a priori* del árbol, *P(D|T)* es la verosimilitud (también conocida como *likelihood*) y *P(D)* es la probabilidad del alineamiento. Para fines prácticos, *P(D)* constituye la sumatoria del numerador *P(T)P(D|T)* sobre todas las posibles topologías *T*:

$$P(T|D) = \frac{P(T)P(D|T)}{\sum_T P(T)P(D|T)}. \quad (2)$$

En otras palabras, al sumar *P(T)P(D|T)* para todos los *T* posibles obtenemos *P(D)*.

Aplicación del teorema de Bayes en filogenética

En la práctica no es posible calcular el denominador de la ecuación (2), ya que la cantidad de topologías posibles (la manera en que las especies se agrupan) incrementa exponencialmente con el número de especies *n*. Repitiendo el mismo ejemplo de la primera parte, un alineamiento con tres especies tiene tres topologías posibles; cuatro especies, 15 topologías posibles; cinco especies tienen 105 topologías, y si hablamos de un alineamiento de 50 especies –muy común en estudios biológicos– tendríamos $2{,}75\times10^{76}$ topologías posibles. Por tanto, computacionalmente no es realista calcular la verosimilitud de tal cantidad de árboles.

Metropolis-Hastings

Para solucionar este problema se usa el algoritmo de Metropolis-Hastings (Metropolis *et al.*, 1953; Hastings, 1970), el cual se basa en una cadena markoviana de Monte Carlo (MCMC, por sus siglas en inglés). Dicho algoritmo explora en el espacio de topologías y toma una muestra representativa de la distribución *P(T|D)*, que es la distribución posterior de la cual queremos obtener *T*. En otras palabras, con Metropolis-Hastings se

analizan individualmente muchas topologías posibles y se toma una muestra de ellas; sin embargo, esta muestra no es aleatoria, más bien representativa, de la distribución posterior de topologías *P(T|D)*. Los pasos del algoritmo de Metropolis-Hastings son los siguientes:

1. Establecer un árbol inicial T_i, el cual constituye la topología "actual".
2. Modificar ligeramente la topología T_i y llamarla T_j (T_j ahora constituye el árbol "candidato").
3. Calcular la relación *A* entre las probabilidades posteriores de las topologías actual T_i y candidata T_j

$$A = \frac{P(T_j|D)}{P(T_i|D)}. \tag{3}$$

4. Si $A > 1$ aceptar T_j como el nuevo árbol actual. En caso contrario, aceptar T_j con probabilidad *A* y constituirlo en el nuevo T_i.
5. Volver al paso 2.

Está demostrado que repetir este algoritmo muchas veces resulta en una muestra significativa de árboles pertenecientes a *P(T|D)* (Metropolis *et al.*, 1953; Hastings, 1970). Ahora, aplicando el teorema de Bayes (1), la ecuación (3) se puede reescribir como:

$$A = \frac{\frac{P(T_j)P(D|T_j)}{P(D)}}{\frac{P(T_i)P(D|T_i)}{P(D)}}, \tag{4}$$

y simplificando los denominadores:

$$A = \frac{P(T_j)P(D|T_j)}{P(T_i)P(D|T_i)}. \tag{5}$$

Por ende, el tomar una muestra representativa de *T* a partir de la distribución objetivo *P(T|D)* se reduce a poder calcular la verosimilitud de topologías actuales $P(D|T_i)$ y candidatas $P(D|T_j)$, y de asumir probabilidades *a priori* para cada una de dichas topologías. En muchos casos se asume que la probabilidad *a priori* de cada topología es la misma, por lo que éstas también se simplificarían. No obstante, es igualmente posible asignar probabilidades *a priori* para las longitudes de rama de *T* a partir de una distribución exponencial, tal es el caso del programa MrBayes (Huelsenbeck & Ronquist, 2001). En cuanto a la verosimilitud *P(D|T)*, el cálculo se efectúa utilizando el algoritmo "prunning" (Felsenstein, 1973).

Pasos generales para inferir una filogenia bajo un modelo bayesiano

Como se estableció anteriormente, la inferencia bayesiana de filogenias requiere también de estimar verosimilitudes de distintas topologías. Combinando lo que se describió en la primera parte de esta revisión con los métodos descritos aquí, los pasos generales para la inferencia bayesiana de filogenias son los siguientes:

1. Proponer una topología inicial.
2. A partir del alineamiento observado *D* y la topología propuesta en el paso 1, calcular la verosimilitud para cada posición (o columna) de *D* utilizando el algoritmo "prunning". Para conocer las probabilidades de sustitución nucleotídica utilizadas en dicho algoritmo, referirse a la sección "Modelos de mutación de ADN" en la primera parte de esta revisión.
3. Una vez obtenida la verosimilitud en cada posición de *D*, calcular la verosimilitud total por medio de la ecuación (2) de la primera parte de esta revisión.
4. Teniendo la verosimilitud de *T*, proponer una topología candidata (similar a la topología actual), calcular su verosimilitud de forma similar siguiendo los pasos 2 y 3, y calcular la relación *A* con la ecuación (5).
5. Si $A > 1$, aceptar la topología candidata como nuevo árbol actual. Caso contrario, aceptar la topología candidata con probabilidad *A*.
6. Repetir los pasos 2 a 5 hasta haber obtenido una muestra representativa de árboles de *P(T|D)*.

Generación de la topología final a partir de la muestra de *P(T|D)*

Es importante describir un paso más para finalizar la reconstrucción de una filogenia con el método bayesiano descrito aquí. Hasta ahora hemos logrado una muestra representativa de árboles de la distribución *P(T|D)*, pero ¿cuál de todas esas topologías se reporta al final? Una forma de abordar este problema consiste en estimar distancias entre todos los árboles de la muestra y tomar como representante al que se encuentre al medio (Li *et al*., 2000; Critchlow, Pearl & Qian, 1996). Otra posibilidad consiste en observar la frecuencia de cada clado en la muestra total de topologías y reportar todas las especies en los clados donde estén más frecuentes (Huelsenbeck, Ronquist, Nielsen & Bollback, 2001; Larget & Simon, 1999).

Finalmente, en cuanto a las especificaciones del Metropolis-Hastings MCMC, es conveniente prestar atención a la frecuencia con que se toman las muestras de *P(T|D)*. Como primer punto, es bueno descartar la primera parte de muestras, ya que no todas ellas pertenecerán a *P(T|D)* (éstas corresponden al *burn-in*). En segundo lugar, no conviene mantener a todas las topologías candidatas aceptadas, ya que en muchos casos serán muy parecidas; es mejor guardar los árboles cada cierto número de repeticiones del algoritmo de Metropolis-Hastings, para así obtener una muestra más representativa de *P(T|D)* (Huelsenbeck *et al*., 2001; Felsenstein, 2004). Es importante notar que para la inferencia bayesiana de filogenias no es necesario utilizar *bootstrapping* como medida de incertidumbre, ya que la probabilidad posterior *P(T|D)* cumple con esta función. La filogenia final reportada contiene probabilidades posteriores para cada clado de la filogenia y cada uno de estos valores describe la probabilidad del clado en cuestión.

Ejemplo de algoritmo para inferencia bayesiana

Enseguida se desarrolla un ejemplo muy sencillo para inferir una filogenia usando el algoritmo de Metropolis-Hastings para inferencia bayesiana. Al igual que en la primera parte de esta revisión, utilizaremos el alineamiento del ejemplo 1.1.1 del artículo "Modelo de estimación de pesos de árbol filogenético para un cuartet, aplicando conjugación de Hadamard" (publicado también en este número). Dicho alineamiento contiene cuatro especies y 16 posiciones. Convertimos dicho alineamiento a formato FASTA y lo guardamos en un archivo denominado "alineamiento.fas":

```
>E1
CCATCAAACGTGTGAC
```

```
>E2
ACAGCAATGTTATCTC
>E3
CCATTGAAGATGCGTT
>E4
ACAGTAGTGTTACCAG
```

Posteriormente, consideramos posibles topologías para las especies E1, E2, E3 y E4. En total existen 15 posibles topologías, las cuales las escribimos en formato NEWICK y las guardamos en un archivo denominado "topologias.tre" (visualizadas en la Figura 1):

```
( ( E1 : 1, E2 : 1 ) : 1, ( E3 : 1, E4 : 1 ) : 1 ) : 1;
( ( E1 : 1, E4 : 1 ) : 1, ( E2 : 1, E3 : 1 ) : 1 ) : 1;
( ( ( E1 : 1, E2 : 1) : 1, E3 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E1 : 1 , E2 : 1 ) : 1, E4 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E1 : 1, E3 : 1 ) : 1, E2 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E1 : 1, E3 : 1 ) : 1, E4 : 1 ) : 1, E2 : 1 ) : 1;
( ( E1 : 1, E3 : 1 ) : 1, ( E2 : 1, E4 : 1 ) : 1 ) : 1;
( ( ( E1 : 1, E4 : 1 ) : 1, E2 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E1 : 1, E4 : 1 ) : 1, E3 : 1 ) : 1, E2 : 1 ) : 1;
( ( ( E2 : 1, E3 : 1 ) : 1, E1 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E2 : 1, E3 : 1 ) : 1, E4 : 1 ) : 1, E1 : 1 ) : 1;
( ( ( E2 : 1, E4 : 1 ) : 1, E1 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E2 : 1, E4 : 1 ) : 1, E3 : 1 ) : 1, E1 : 1 ) : 1;
( ( ( E3 : 1, E4 : 1 ) : 1, E2 : 1 ) : 1, E1 : 1 ) : 1;
( ( ( E3 : 1, E4 : 1 ) : 1, E1 : 1 ) : 1, E2 : 1 ) : 1;
```

Finalmente, usando los archivos que se acaban de crear ("alineamiento.fas" y "topologias.tre") como input, desarrollo a continuación un programa corto (escrito en el lenguaje de programación R) a fin de ejemplificar el algoritmo Metropolis-Hastings para la inferencia filogenética bayesiana (nótese que los pasos del algoritmo también están mencionados en el programa):

```
#######-INICIO DEL PROGRAMA-#######

library(ape)
library(phangorn)

#Alineamiento de ADN.
D <- phyDat(read.FASTA("alineamiento.fas")) #Posibles topologias.
T <- read.tree("topologias.tre")
```

```
##-------Metropolis-Hastings para inferencia bayesiana-------## cat("\nEjemplo Metropolis-Hastings para inferencia Bayesiana\n")
#Se comienza con la topología inicial (paso 1).
indice_actual <- 1
T_actual <- T[ [ indice_actual ] ]

for (i in 2:length(T) ) {
        #La topología candidata es la siguiente en la lista de T (paso 2). T_candidata <- T[[i]]

        #Se calcula la verosimilitud de la topología actual.
        logV_actual <- pml(T_actual,D)

        #Se calcula la verosimilitud de la topología candidata.
        logV_candidata <- pml(T_candidata,D)

        #Se calcula la relacion de verosimilitudes A (paso 3).
        A <- logV_actual$logLik/logV_candidata$logLik

        if (A>1) { #Si A>1 (paso 4 parte 1).
                T_actual <- T_candidata
                indice_actual <- i
                print(paste("T candidata =",i,", log V = ",logV_candidata$logLik))
        } else {
                #Si A<1 (paso 4 parte 2).
                nuevoIndice <- sample(c(indice_actual,i),1,prob=c(A,1-A)) T_actual <- T[[nuevoIndice]]

                if (nuevoIndice==indice_actual) {
                        print(paste("T actual =",indice_actual,", log V = ",logV_actual$logLik))
                } else {
                        print(paste("T candidata =",indice_actual,", log V = ",logV_candidata$logLik))
                }
                        indice_actual <- nuevoIndice
        }
}

#######-FIN DEL PROGRAMA-#######
```

El output de este programa es el siguiente:

```
Ejemplo Metropolis-Hastings para inferencia Bayesiana
[1]  "T actual = 1 , log Verosimilitud = -87.1232112880638"
[1]  "T candidata = 3 , log Verosimilitud = -87.0763445725479"
```

[1] "T candidata = 4 , log Verosimilitud = -87.0328449489164"
[1] "T candidata = 5 , log Verosimilitud = -86.6093614372388"
[1] "T actual = 5 , log Verosimilitud = -86.6093614372388"
[1] "T candidata = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"

Como se puede corroborar, el algoritmo de Metropolis-Hastings converge en la topología 7; sin embargo, cabe recordar que aquí usamos Metropolis-Hastings solamente para desarrollar un ejemplo ilustrativo. En realidad, para un alineamiento tan reducido serán suficientes otros métodos basados en distancias genéticas para inferir la filogenia. La máxima verosimilitud y la inferencia bayesiana usando Metropolis-Hastings MCMC son más útiles para un mayor número de especies en alineamientos más grandes. En tales casos, no existirá solamente una topología con la mayor probabilidad posterior, sino varias, por lo cual se deberán utilizar los métodos descritos en la sección "Generación de la topología final a partir de la muestra de *P(T|D)*" para obtener el resultado final. Finalmente, dependiendo de la cantidad de especies y de la longitud del alineamiento se prefiere un método sobre otro (e. g. Inferencia bayesiana sobre máxima verosimilitud, o viceversa). Si se desea revisar más detalladamente los factores que influyen en la selección de un método de inferencia filogenética, remitirse a Peña (2011).

Software para inferencia bayesiana de filogenias

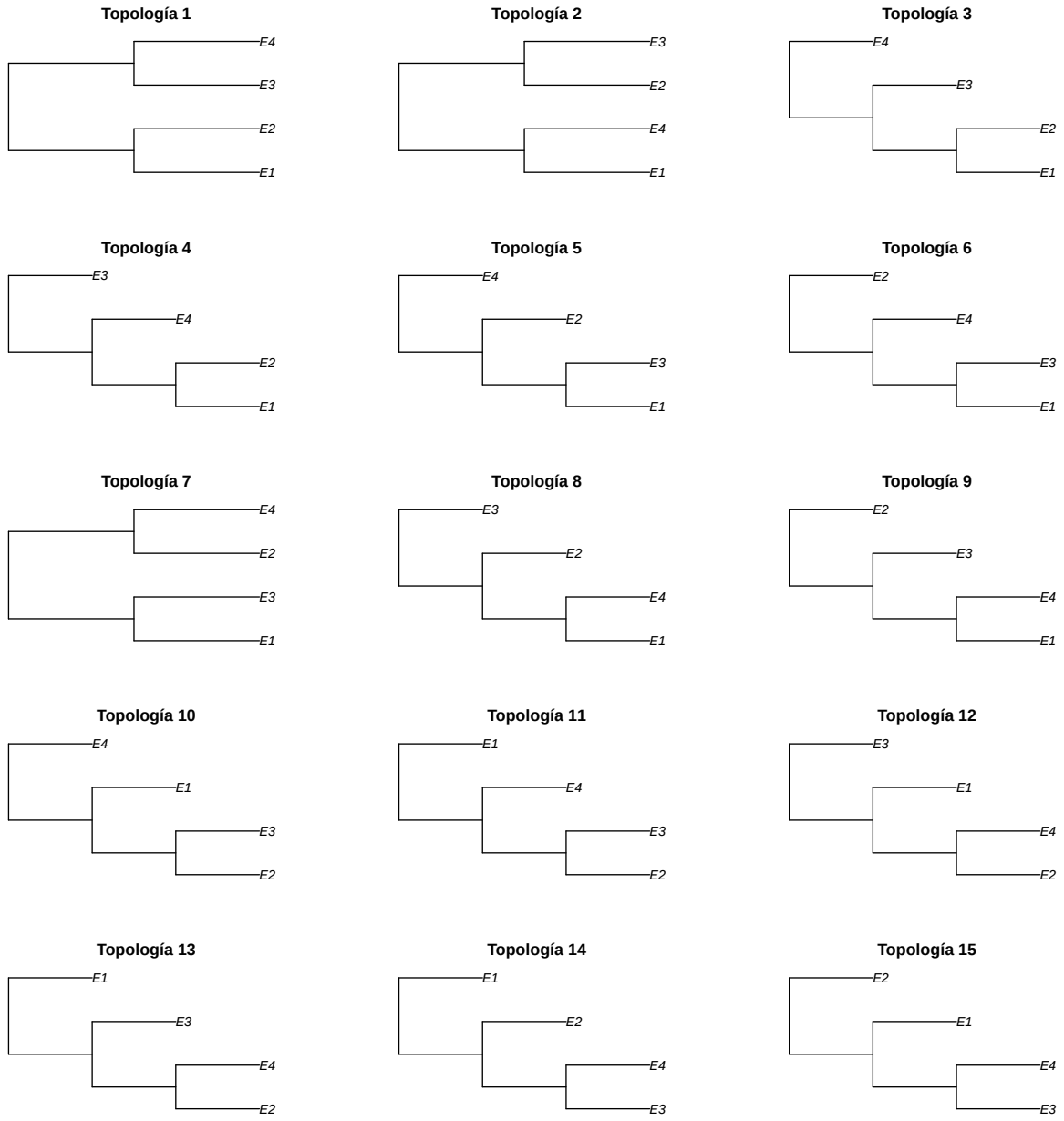
Existen muchos programas que hacen inferencia bayesiana de filogenias. Aquí se nombran los más usados en la actualidad:

MrBayes. Este programa fue desarrollado inicialmente por Huelsenbeck & Ronquist (2001), con una nueva versión publicada más recientemente (Ronquist *et al.*, 2012). Este es el programa clásico empleado en reconstrucción filogenética con un método bayesiano. Utiliza todos los elementos descritos aquí: probabilidades *a priori*, modelos de mutación de ADN y Metropolis-Hastings MCMC para encontrar el árbol con la mayor probabilidad posterior.

BEAST. Este programa fue originalmente introducido por Drummond & Rambaut (2007). Además de encontrar el árbol con la mayor probabilidad posterior, BEAST incluye rutinas que calculan un reloj molecular (o datación de filogenias) bajo distintos modelos.

RevBayes. Desarrollado por Höhna *et al.* (2016), sus funciones comprenden inferencia filogenética, MCMC, relojes moleculares, selección de modelos, estimación de tasas de diversificación, etcétera. Este programa incluye su propio intérprete, lo que lo hace particularmente útil a la hora de desarrollar tareas más complejas.

Figura 1.
Topologías posibles para el alineamiento *D* del ejemplo en la sección "Ejemplo de algoritmo para inferencia bayesiana". La topología con la máxima verosimilitud es la 7.



Referencias

Critchlow, D. E., Pearl, D. K. & Qian, C. (1996). The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology,* 45, 323-334. doi: 10.1093/sysbio/45.3.323

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology,* 7, 214. doi: 10.1186/1471-2148-7-214

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology,* 22, 240-249. doi: 10.1093/sysbio/22.3.240

Felsenstein, J. (2004). *Inferring phylogenies*, vol. 2. Sunderland, Massachusetts: Sinauer associates.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109. doi: 10.1093/biomet/57.1.97

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P. & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology,* 65, 726-736. doi: 10.1093/sysbio/syw021

Huelsenbeck, J. P. & Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17, 754-755. doi: 10.1093/bioinformatics/17.8.754

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294, 2310-2314. doi: 10.1126/science.1065889

Larget, B. & Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16, 750-759. doi: 10.1093/oxfordjournals.molbev.a026160

Li, S., Pearl, D. K. & Doss, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American statistical Association*, 95, 493-508. doi: 10.1080/01621459.2000.10474227

Mau, B. & Newton, M. A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 6, 122-131. doi: 10.1080/10618600.1997.10474731

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,* 21, 1087-1092. doi: 10.1063/1.1699114

Peña, C. (2011). Métodos de inferencia filogenética. *Revista Peruana de Biología* 18, 265-267.

Rannala, B. & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43, 304-311. doi: 10.1007/BF02338839

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61, 539-542. doi: 10.1093/sysbio/sys029

Yang, Z. & Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution,* 14, 717-724. doi: 10.1093/oxfordjournals.molbev.a025811