Tequio *4*(11), 2021: 69-79
ISSN: 2594-0546

Métodos de reconstrucción filogenética I: máxima verosimilitud

Methods for phylogenetic reconstruction I: maximum likelihood

Pablo Duchen[1*]

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

Resumen - La inferencia filogenética es ampliamente utilizada en biología evolutiva, la cual tiene el objetivo de encontrar las relaciones evolutivas entre diferentes especies y representarlas en la forma de un árbol filogenético (o filogenia). Existen varios métodos estadísticos para la inferencia filogenética. En esta revisión se presenta la máxima verosimilitud como modelo de reconstrucción filogenética, método que consiste en calcular la verosimilitud de múltiples filogenias candidatas y reportar aquella con el valor máximo como la filogenia representativa de un grupo de organismos. En la presente revisión se explica cómo se calcula la verosimilitud de una filogenia a partir de secuencias de ADN provenientes de varias especies. También se presentan modelos de mutación de ADN para calcular probabilidades de transición entre nucleótidos, los cuales son usados en la estimación de la verosimilitud. Se muestra también un ejemplo ilustrativo sencillo que aplica los pasos necesarios para inferir una filogenia y se explica el software más usado para inferencia bajo máxima verosimilitud para alineamientos de ADN más grandes.

Palabras clave: Prunning, Jukes-Cantor, inferencia filogenética, alineamiento, bootstrap, modelos de mutación de ADN.

Abstract - Phylogenetic inference is widely used in evolutionary biology, aiming to find evolutionary relationships between different species and report the result in the form of a phylogenetic tree (phylogeny). There are several statistical methods used for phylogenetic inference. In this review, the method of maximum likelihood for phylogenetic reconstruction is presented. This technique consists of finding the likelihood of multiple candidate phylogenies, and report the one with the highest likelihood as a representative of the evolutionary relationships of a group of species. In this paper, the likelihood calculation of a phylogeny from multiple-species DNA sequences is reviewed. Also, some key DNA mutation models to calculate transition probabilities between nucleotides are presented. Such transition probabilities are used in the likelihood calculation of a given phylogeny. A simple example is shown to illustrate the necessary steps to infer a phylogeny, as well as the most common software for maximum likelihood inference for larger DNA alignments.

Keywords: Prunning, Jukes-Cantor, phylogenetic inference, alignment, bootstrap, DNA mutation models.

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

Introducción

La sistemática filogenética tiene como objetivo encontrar las relaciones evolutivas o de parentesco entre diferentes especies o distintos taxones supraespecíficos. Dichas conexiones evolutivas se representan comúnmente en la forma de un árbol filogenético, donde organismos emparentados están unidos por medio de líneas que simbolizan las ramas del árbol, y donde los nodos significan ancestros comunes entre especies o clados. Dada la relevancia biológica de conocer las relaciones evolutivas entre distintas especies, inferir una filogenia a partir de datos morfológicos o moleculares es una práctica muy efectuada en biología evolutiva y taxonomía. Por otro lado, las filogenias también se usan para reconstruir caracteres ancestrales (Pagel, 1999), establecer relojes moleculares (Bronham & Penny, 2003), o para estudiar la evolución de caracteres morfológicos (Duchen, Alfaro, Rolland, Salamin & Silvestro, 2020).

Existen varios métodos estadísticos para la inferencia o reconstrucción filogenética (Brocchieri, 2001); entre ellos, la máxima verosimilitud (ML) y la inferencia bayesiana son probablemente los más usados en la actualidad. A diferencia de procedimientos basados en distancias genéticas entre secuencias (como el *neighbor joining* o el UPGMA), o basados en máxima parsimonia (Edwards & Cavalli-Sforza, 1963; Peña, 2011), ML y el bayesiano utilizan las verosimilitudes de cada posición (o columna) en un alineamiento de secuencias de ADN para inferir una filogenia. En este trabajo se van a desarrollar primero los pasos necesarios para calcular una filogenia basada en ML y en el siguiente artículo se plantearán los pasos para realizar una inferencia bayesiana. Se comenzará con el algoritmo necesario para calcular la verosimilitud de un alineamiento (dada una filogenia particular); luego se describirán dos ejemplos de modelos de mutación de ADN, los cuales son usados en el cálculo de verosimilitudes; se concluye con un ejemplo simple de inferencia filogenética, mencionando además el software usado para alineamientos más grandes.

Cálculo de la verosimilitud de una topología

Es importante comenzar 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 la filogenia, árbol o topología T que mejor describa dicho alineamiento. A lo largo de este documento se usarán los vocablos filogenia y árbol indistintamente para referirse a T, al igual que los términos alineamiento o datos para referirse a D.

Dado un alineamiento D y asumiendo que la evolución en cada posición de D y cada rama de T es independiente, la verosimilitud está determinada por:

$$P(D|T) = \prod_{k=1}^{m} P\left(D^{(k)}|T\right), \quad (1)$$

Donde $D^{(k)}$ corresponde al alineamiento en la posición k (con un total de m posiciones). Esto significa que si podemos calcular la verosimilitud en cada posición del alineamiento separadamente, la verosimilitud total será simplemente el producto de todas las posiciones de alineamiento. Para fines prácticos y evitar problemas numéricos es mejor utilizar la versión logarítmica de dicha ecuación:

De esta manera, en vez de multiplicar todas las verosimilitudes, se suman los logaritmos de $P(D^{(k)}|T)$ para cada posición. El problema con multiplicar verosimilitudes es que éstas representan probabilidades con valores entre 0 y 1. Al multiplicar valores menores a 1 varias veces, los ceros decimales aumentan y se pierden rápidamente las cifras no periódicas durante la multiplicación. Por este motivo se debe usar la versión logarítmica.

Para calcular la verosimilitud en una posición se emplea frecuentemente el método o algoritmo de "pruning" de Felsenstein (1973), el cual se describirá a continuación.

Algoritmo "prunning"

Este algoritmo es muy eficiente para calcular la verosimilitud de una filogenia y está basado en verosimilitudes condicionales para cada clado de *T*. Se llama "condicional" a dicha verosimilitud porque su valor depende de los nucleótidos que estén en el extremo de cada clado. Aquí se denomina $V^{(k)}$ a la verosimilitud condicional de cada clado en una filogenia en la posición *k* de un alineamiento. Para no sobrecargar la notación vamos a dejar momentáneamente de lado la indicación de la posición *k*; por ejemplo, el árbol de la Figura 1 tiene las especies E_1, E_2 y E_3, las cuales, en esa posición del alineamiento, muestran las bases A, G y G, respectivamente. Por el contrario, las bases de las especies ancestrales E_{12} y E_{123} se desconocen y pueden tomar cualquier valor entre los nucleótidos A, C, G, T.

Las bases de ADN en los extremos del árbol son datos observados, por tanto, los valores de *V* en los extremos serían: V_{E1}(A,C,G,T)= (1,0,0,0), ya que se observa al nucleótido A en la especie E_1; V_{E2}(A,C,G,T)= (0,0,1,0), porque se observa al nucleótido G en la especie E_2; y V_{E3}(A,C,G,T)= (0,0,1,0). Una vez calculadas las *V*s en los extremos del árbol, se van a calcular las *V*s en los nodos internos.

La verosimilitud condicional del nodo interno E_{12} está dada por dos posibilidades: la de cambiar del estado de E_{12} a A, y de E_{12} a G, a lo largo de la rama t_{12} que separa a ambos nodos. Por tanto,

$$V_{\mathrm{E}_{12}} = (P(A|\mathrm{E}_{12}, t_{12}) \times 1)(P(G|\mathrm{E}_{12}, t_{12}) \times 1).$$

En esta ecuación, el primer factor corresponde a la probabilidad de cambiar del estado de E_{12} a A, y el segundo factor a la probabilidad de cambiar del estado de E_{12} a G. Nótese aquí que multiplicamos ambos factores por 1, que corresponden a las probabilidades de observar las bases A y G en las puntas de dicho árbol, respectivamente (estas probabilidades corresponden a V_{E1} y V_{E2}, calculados anteriormente).

Se procede ahora a calcular la verosimilitud condicional en la raíz del árbol. Si se define *x* como el nucleótido que corresponde a E_{12}, entonces la verosimilitud condicional en la raíz del árbol de la Figura 1 está dada por:

$$V_{\mathrm{E}_{12}} = \left(\sum_x P(x|\mathrm{E}_{123}, t_{123} - t_{12}) V_{\mathrm{E}_{12}}\right) P(G|\mathrm{E}_{123}, t_{123}) \times 1.$$

El primer factor muestra la probabilidad de cambiar del estado de E_{123} al nucleótido *x*, a lo largo de la rama de longitud t_{123}-t_{12} dada la probabilidad V_{E12} (calculada en el paso anterior). Nótese que aquí se suman todas las probabilidades de *x* ($x \in$ {A,C,G,T}), ya que se desconoce el nucleótido correspondiente a E_{12}. El segundo factor atañe a la probabilidad de cambiar del estado de E_{123} a G a lo largo de la rama t_{123}. Dicho factor igualmente se multiplica por 1, que refiere a la probabilidad de observar una G en el extremo del árbol.

En general, y para cualquier filogenia, asumiendo que el clado en cuestión tiene como nucleótido *s* ($s \in$ {A,C,G,T}) y tiene dos descendientes con nucleótidos *x* ($x \in$ {A,C,G,T}) y *y* ($y \in$ {A,C,G,T}), con longitudes de rama t_x y t_y, entonces cada V_E está dada por:

$$V_E = \left(\sum_x P(x|s,t_x)V_x\right)\left(\sum_y P(y|s,t_y)V_y\right)$$

De esta manera, comenzando por la punta del árbol, se calculan las verosimilitudes condicionales descendiendo por cada nodo hasta llegar a la raíz. Al final, la verosimilitud total de la filogenia *T* en la posición *k* (retomando la notación original) es:

$$V^{(k)} = P\left(D^{(k)}\middle|T\right) = \sum_x \pi_x V_{E_{raíz}}^{(k)}(x),$$

donde π_x es la probabilidad *a priori* del nucleótido *x* –la cual se puede estimar por su frecuencia en el alineamiento– y $E_{raíz}$ es el nucleótido en la raíz del árbol, correspondiente a E_{123} en el ejemplo de la Figura 1. Finalmente, todas las probabilidades *P(x|s,t)* o *P(y|s,t)* se calculan con diversos modelos de mutación de ADN, los cuales se describen en la sección "Modelos de mutación de ADN".

Inferencia filogenética

La inferencia por ML funciona de la siguiente manera: dadas las topologías candidatas para un alineamiento *D* particular, se pueden calcular las verosimilitudes de cada una (utilizando el algoritmo "prunning"). Luego, la topología con la mayor verosimilitud será la filogenia correspondiente a *D*. Hay dos aspectos importantes para tomar en cuenta al realizar la inferencia por ML: el *bootstrapping* y la búsqueda de topologías.

Bootstrapping. En inferencia por ML se recurre al *bootstrap* para obtener una medida de incertidumbre para el árbol con la máxima verosimilitud. El *bootstrap* en filogenética consiste en: 1) tomar muestras con reemplazo de las columnas de un alineamiento, 2) formar un nuevo alineamiento con dichas columnas, y 3) volver a inferir la topología con el nuevo alineamiento. Para ser más preciso, si *D* tiene *m* columnas, entonces se toman *m* muestras con reemplazo y se infiere la filogenia. Este proceso se repite múltiples veces. Estudios que utilizan hasta 2 500 secuencias muestran que 100 a 500 repeticiones de *bootstrap* son suficientes, pero para criterios más conservadores se llegan a hacer varios miles de repeticiones (Pattengale, Alipour, Bininda-Emonds, Moret & Stamatakis, 2010). En la filogenia final (aquella con la máxima verosimilitud) se reporta el porcentaje de ocasiones que cada clado se mantiene en las repeticiones del *bootstrapping*. Clados con valores de *bootstrap* mayores a 75% se consideran con buen soporte estadístico.

Búsqueda de topologías. Otro aspecto importante constituye la búsqueda de topologías. Para alineamientos con pocas especies es posible calcular la verosimilitud de todas las topologías posibles, lo que se conoce como una búsqueda exhaustiva. Sin embargo, para alineamientos con muchas especies la cantidad de topologías para analizar es muy grande, por lo que se emplea la búsqueda heurística (aproximada). Para dar un ejemplo, un alineamiento con tres especies tiene tres topologías posibles; cuatro especies, 15 topologías posibles (Fig. 2);

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. Para solucionar esto, existen algoritmos que hacen una búsqueda de topologías solamente entre aquellas con mayor verosimilitud (Stamatakis, 2014).

Modelos de mutación de ADN

Se continúa ahora con la descripción de modelos de mutación de ADN, los cuales se usan para calcular verosimilitudes. Existen distintos modelos para estimar la probabilidad de cambiar de un nucleótido a otro a lo largo de una rama. La complejidad de dichos modelos radica en la reversibilidad de la sustitución de nucleótidos, o en la inclusión de eventos denominados "transiciones" y "transversiones". Desde el punto de vista genético-molecular, una transición es una sustitución entre purinas (nucleótidos A y G) o pirimidinas (C y T), mientras que una transversión es un cambio de una purina a una pirimidina, o viceversa.

Jukes-Cantor

El modelo más simple es el de Jukes-Cantor (Jukes & Cantor, 1969), donde se asume que la probabilidad de que un nucleótido modifique a los otros tres es la misma. Partiendo de que la tasa instantánea de cambio de un nucleótido específico es $\mu/3$, y que la frecuencia de cada nucleótido es 1/4 ($\pi_A = \pi_C = \pi_G = \pi_T = ¼$), entonces la tasa de sustitución total de un nucleótido cualquiera es $\mu/3 + \mu/3 + \mu/3 + \mu/3 = 4\mu/3$.

Para calcular la probabilidad de un evento de sustitución a lo largo de una rama de longitud t se emplea la distribución de Poisson. En este caso, la probabilidad de una "no" sustitución es $e^{-4\mu t/3}$, donde $4\mu/3$ corresponde a la tasa de sustitución calculada en el párrafo anterior y la probabilidad de al menos un evento de sustitución es $1\text{-}e^{-4\mu t/3}$. Por tanto, la probabilidad de cambio de un nucleótido s a otro x a lo largo de una rama de longitud t viene dada por:

$$P(x|s,\mu,t) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}\mu t}\right),$$

donde el factor 1/4 indica la probabilidad de que el último evento de sustitución resulte en el nucleótido x.

El modelo de Kimura (1980) también asume que las frecuencias nucleotídicas π_X ($X \in \{A,C,G,T\}$) son las mismas. Sin embargo, a diferencia del modelo de Jukes-Cantor, el de Kimura distingue entre transiciones y transversiones. No se desarrollará aquí este modelo, pero se presentará uno más general que incluye al de Kimura, además de otros como casos especiales.

Tamura-Nei

Este modelo descrito originalmente por Tamura & Nei (1993) hace diferencia entre transiciones y transversiones, y también permite que las frecuencias nucleotídicas π_X sean distintas entre sí. Para tomar en cuenta transiciones y transversiones –y siguiendo la notación en Felsenstein (2004)– se definirá como α_R a la probabilidad de escoger una purina a partir una purina, y α_Y a la probabilidad de escoger una pirimidina a partir de una pirimidina. Además, se definirá como β a la probabilidad de elegir un nucleótido cualquiera a partir de las cuatro bases A, C, G y T.

Con base en estas definiciones la tasa instantánea de cambio entre las purinas e.g. G a A sería $\alpha_R \frac{\pi_A}{\pi_A+\pi_G} + \beta\pi_A$. Aquí, el primer término representa la probabilidad de escoger otra purina (A) dado que se parte de una purina (G), mientras que el segundo término representa la probabilidad de escoger al nucleótido A a partir de

cualquier otro. Para el caso de una transversión (por ejemplo de C a A) la tasa instantánea de cambio es simplemente $\beta\pi_A$. De acuerdo con estos ejemplos, es sencillo escribir las tasas de cambio para el resto de nucleótidos.

Para calcular la probabilidad de un evento de sustitución a lo largo de una rama de longitud t también se utilizará la distribución de Poisson. Si se comienza con una purina, la probabilidad de que no haya ningún evento de transición a lo largo de t es $e^{-\alpha_R t}$, la probabilidad de transversiones es $(1-e^{-\beta t})$, la de transiciones pero no transversiones es $(1-e^{-\alpha_R t})e^{-\beta t}$, la de que no ocurra ningún evento es $e^{-(\alpha_R+\beta)t}$, etcétera. Estos mismos criterios para cálculo de probabilidades aplican para las pirimidinas. Ahora sí, la probabilidad de sustitución total entre e.g. A y G a lo largo de t es:

$$P(G|A,t) = e^{-\beta t}(1-e^{-\alpha_R t})\frac{\pi_G}{\pi_A+\pi_G} + \left(1-e^{-\beta t}\right)\pi_G.$$

En otras palabras, para obtener una G a partir de una A existen dos posibilidades: 1) no puede haber ninguna transversión $(e^{-\beta t})$ y tiene que haber una transición $(1-e^{-\alpha_R t})$ que resulte en una G, o 2) pueden existir transversiones $(1-e^{-\beta t})$ siempre y cuando la última resulte en una G. De esta misma manera pueden escribirse expresiones para las otras sustituciones entre los distintos pares de nucleótidos.

Se finaliza esta sección mencionando los modelos F84 (Felsenstein & Churchill, 1996), HKY (Hasegawa, Kishino & Yano, 1985) y el modelo *General time reversible* (GTR) (Lanave, Preparata, Sacone & Serio, 1984; Tavaré, 1986). En el modelo F84 $\alpha_R = \alpha_Y$. Por otro lado, si $\frac{\alpha_R}{\alpha_Y} = \frac{\pi_A+\pi_C}{\pi_G+\pi_T}$ entonces obtenemos el modelo HKY. Finalmente, el modelo GTR generaliza los modelos vistos anteriormente, ya que incluye seis tasas de sustitución (una para cada par de nucleótidos). Si bien no es necesaria su descripción a detalle en la presente revisión, este modelo está implementado en los programas clásicos de inferencia filogenética (ver sección "Software para inferencia filogenética con ML").

Ejemplo de algoritmo para inferencia por ML

Ahora se va a desarrollar un ejemplo muy sencillo para inferir una filogenia usando ML. Se utilizará 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 de la revista). Dicho alineamiento contiene cuatro especies y 16 posiciones; se convierte a formato FASTA y se guarda en un archivo denominado "alineamiento.fas":

```
>E1
CCATCAAACGTGTGAC
>E2
ACAGCAATGTTATCTC
>E3
CCATTGAAGATGCGTT
>E4
ACAGTAGTGTTACCAG
```

Posteriormente, se consideran posibles topologías para las especies E1, E2, E3 y E4. En total existen 15 posibles topologías, las cuales se pueden escribir en formato NEWICK y se guardan en un archivo denominado "topologias.tre" (visualizadas en la Figura 2):

```
( ( 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, utilizando los archivos que recién se crearon ("alineamiento.fas" y "topologias.tre") como input, se desarrolla a continuación un programa corto (escrito en el lenguaje de programación R) para ejemplificar el algoritmo de inferencia filogenética utilizando ML:

```
#######-INICIO DEL PROGRAMA-#######
library(ape)
library(phangorn)
#Alineamiento de ADN.
D <- phyDat(read.FASTA("alineamiento.fas")) #Posibles topologias.
T <- read.tree("topologias.tre")

##-------Maxima Verosimilitud-------##
cat ("\nEjemplo de inferencia con ML\n")

#El vector logV guardara las verosimilitudes de cada topología.
logV <- numeric(length(T))

#La función pml de la librería phangorn calcula la verosimilitud.
for (i in 1:length(T)) {
        logV[i] <- pml(T[[i]],D)$logLik
        print(paste("Log Verosimilitud topologia",i,"=",logV[i] ) )
}
```

```
#Aqui se escoge la topología con la máxima verosimilitud. print ( paste ( "Topologia con la maxima
verosimilitud =",
                    which ( logV==max(logV) ) ) )
#######-FIN DEL PROGRAMA-#######
```

El output de este programa es el siguiente:

```
Ejemplo de inferencia con ML
[1] "Log Verosimilitud topologia 1 = -87.1232112880638"
[1] "Log Verosimilitud topologia 2 = -87.8578012541171"
[1] "Log Verosimilitud topologia 3 = -87.0763445725479"
[1] "Log Verosimilitud topologia 4 = -87.0328449489164"
[1] "Log Verosimilitud topologia 5 = -86.6093614372388"
[1] "Log Verosimilitud topologia 6 = -86.7630718337047"
[1] "Log Verosimilitud topologia 7 = -85.6551750661665"
[1] "Log Verosimilitud topologia 8 = -87.61408828079"
[1] "Log Verosimilitud topologia 9 = -87.8112983008873"
[1] "Log Verosimilitud topologia 10 = -87.5197055341976"
[1] "Log Verosimilitud topologia 11 = -87.5197055341976"
[1] "Log Verosimilitud topologia 12 = -86.2876683750476"
[1] "Log Verosimilitud topologia 13 = -86.331167998679"
[1] "Log Verosimilitud topologia 14 = -87.3545380111076"
[1] "Log Verosimilitud topologia 15 = -87.5082484075735"
[1] "Topologia con la maxima verosimilitud = 7"
```

Como se puede corroborar, la topología con la máxima verosimilitud es la 7 (Figura 2). En la realidad, para un alineamiento tan reducido, otros métodos basados en distancias genéticas serán suficientes para inferir la filogenia. ML es más útil para alineamientos con más especies y más posiciones.

Software para inferencia filogenética con ML

Existen muchos programas que calculan filogenias utilizando ML. Históricamente el software PAUP (Swofford, 2002) también incluye un método ML. Sin embargo, se desarrollaron otros programas con mayor velocidad, aquí se nombrarán los más usados en la actualidad.

RAxML. Desarrollado por Stamatakis (2014), este programa toma un alineamiento como input y reporta la filogenia con la máxima verosimilitud. Es muy utilizado debido a su rapidez; Además, RAxML también realiza el *bootstrapping*.

PHYLIP. Desarrollado por Felsenstein y colaboradores (Felsenstein, 2019). La primera versión salió en 1980 y desde entonces fue desarrollándose y actualizándose continuamente. PHYLIP, además de calcular filogenias

empleando ML, también lleva a cabo *bootstrapping* e incluye varios métodos de análisis evolutivo a lo largo de filogenias.

PhyML. Desarrollado por Guindon *et al.* (2010), PhyML consiste en diversos programas que calculan una filogenia utilizando ML. Además, contiene herramientas para la calibración de fósiles en filogenias y para estimar tasas de dispersión.

MEGA. Desarrollado por Hall (2013). También incluye inferencia filogenética con ML, en conjunto con una multitud de herramientas útiles en genética evolutiva.

Figura 1.
Algoritmo "prunning" para calcular la verosimilitud de un árbol *T* en una posición (o columna) *k* del alineamiento *D*. Aquí observamos los nucleótidos A, G y G para las especies E1, E2 y E3, respectivamente. Los nucleótidos en los nodos interiores E_{12} y E_{123} se desconocen, pero eso no impide calcular su verosimilitud, ya que se pueden sumar las cuatro posibilidades correspondientes a los cuatro nucleótidos posibles.

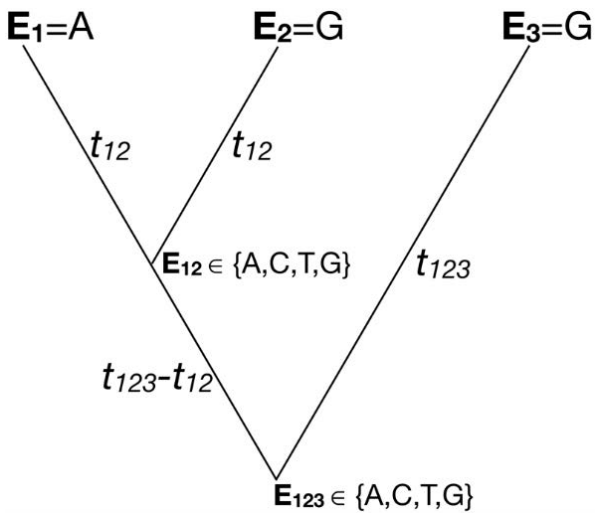
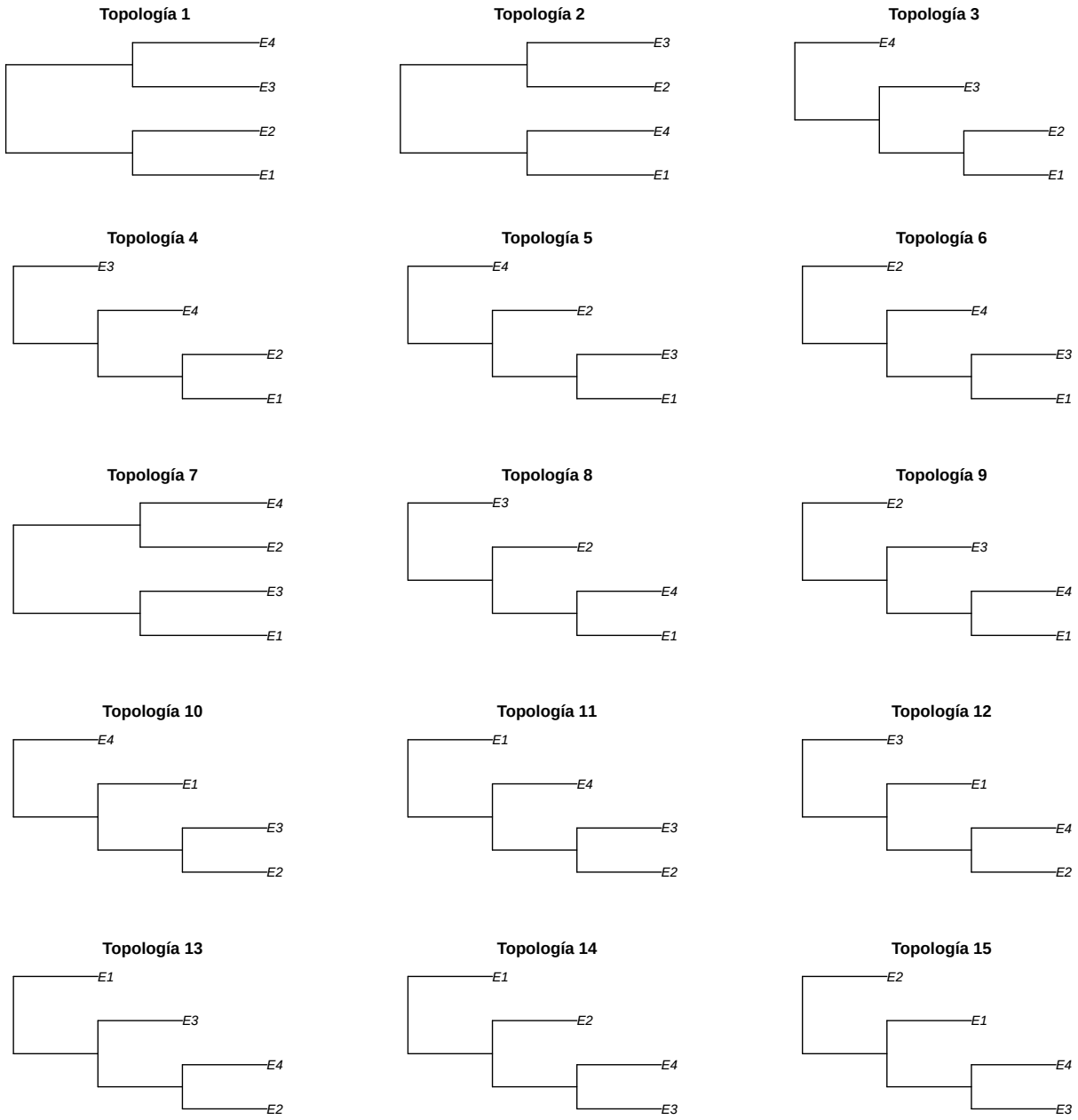


Figura 2.

Topologías posibles para el alineamiento *D* del ejemplo en la sección 4. La topología con la máxima verosimilitud es la 7.



Referencias

Brocchieri, L. (2001). Phylogenetic Inferences from Molecular Sequences: Review and Critique. *Theoretical Population Biology, 59*(1), 27-40. doi: 10.1006/tpbi.2000.1485

Bronham, L. & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics,* 4, 216-224. doi: 10.1038/nrg1020

Duchen, P., Alfaro, M., Rolland, J., Salamin, N. & Silvestro, D. (2020). On the effect of asymmetrical trait inheritance on models of trait evolution. *Systematic Biology* (In press). doi: 10.1093/sysbio/syaa055

Edwards, A. & Cavalli-Sforza, L. (1963). The reconstruction of evolution. *Annals of Human Genetics,* 27, 106-106.

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.

Felsenstein, J. (2019). *PHYLIP (phylogeny inference package) version 3.698*. Recuperado de https://evolution.genetics.washington.edu/phylip.html

Felsenstein, J. & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution,* 13, 93-104. doi: 10.1093/oxfordjournals.molbev.a025575

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* 59, 307-321. doi: 10.1093/sysbio/syq010

Hall, B. G. (2013). Building phylogenetic trees from molecular data with MEGA. *Molecular Biology and Evolution*, 30, 1229-1235. doi: 10.1093/molbev/mst012

Hasegawa, M., Kishino, H. & Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22, 160-174. doi: 10.1007/BF02101694

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. En M. Munro (Ed.), *Mammalian protein metabolism* (pp. 21-132), vol. 3. New Yor: Academic Press.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution,* 16, 111-120. doi: 10.1007/BF01731581

Lanave, C., Preparata, G., Sacone, C. & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution,* 20, 86-93. doi: 10.1007/BF02101990

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature,* 401, 877-884. doi: 10.1093/sysbio/syr124

Pattengale, N., Alipour, M., Bininda-Emonds, O., Moret, B. & Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of Computational Biology,* 17, 337-354. doi: 10.1089/cmb.2009.0179

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

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312-1313. doi: 10.1093/bioinformatics/btu033

Swofford, D. L. (2002). PAUP: phylogenetic analysis using parsimony, version 4.0 b10. Doi: 10.1111/j.0014-3820.2002.tb00191.x

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10, 512-526. doi: 10.1093/oxfordjournals.molbev.a040023

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences,* 17, 57-86.