

# Capítulo 9

## Filogenia y evolución molecular

*Federico Abascal, Iker Irisarri y Rafael Zardoya*

### 9.1. Introducción

Hace aproximadamente 3.500 millones de años surgió la vida en la Tierra. Desde entonces, la vida ha explorado múltiples caminos, unos breves y otros muy extensos. Sólo algunos de ellos han perdurado hasta la actualidad. Es una historia de cambios y adaptación. Una adaptación que se refleja a muchos niveles, desde la maquinaria celular y los genes que gobiernan el desarrollo, hasta modificaciones morfológicas o de comportamiento. Esta historia ha quedado registrada en el genoma de las especies actuales, de modo que el estudio de la información genética nos permite conocer los vericuetos de la historia de la vida, y además nos abre una ventana al laboratorio evolutivo, donde infinidad de opciones han sido probadas, pero sólo una parte de ellas han tenido éxito. Una pregunta clave en Biología es *por qué los seres vivos son como son*, y no se puede responder sin considerar el componente histórico. Los estudios evolutivos intentan resolver ésta y otras cuestiones, y aportan información muy valiosa prácticamente en cualquier nivel de organización en biología. En palabras del célebre biólogo Theodosius Dobzhansky: “*Nada tiene sentido en biología sino es a la luz de la evolución*”.

Cuando hablamos de evolución nos referimos al proceso de descendencia con modificación a partir de un ancestro común [8]. El hecho de que los organismos comparten relaciones ancestro-descendiente nos permite utilizar árboles filogenéticos para representar gráficamente sus relaciones de parentesco. Los árboles filogenéticos tienen dos componentes principales: las relaciones de parentesco (que conocemos como la topología del árbol) y la cantidad de cambio acumulado en cada rama (que conocemos como longitudes de rama).

Los métodos de reconstrucción filogenética permiten hacer uso de la información contenida en el genoma de las especies actuales para *inferir* su historia evolutiva (el árbol filogenético). Hablamos de inferencia porque la reconstrucción filogenética se basa necesariamente en una información incompleta, puesto que sólo tenemos acceso a la información genética de organismos actuales [43]. Estamos, por tanto, ante un problema de encontrar la (o las) hipótesis más probable(s). Por eso, la estadística juega un papel muy importante en el análisis evolutivo.

En origen, los métodos de reconstrucción filogenética fueron concebidos para inferir las relaciones de parentesco entre organismos. Sin embargo, y dado que la evolución subyace a todo proceso biológico, en la actualidad los métodos filogenéticos se utilizan para responder a muchos otros interrogantes. Estos incluyen la reconstrucción de la historia evolutiva de genes, la estimación de tiempos de divergencia, la

reconstrucción de estados ancestrales, el estudio de la selección natural, la detección de recombinación, la dinámica poblacional, teniendo importantes aplicaciones prácticas en epidemiología, criminología y medicina [18].

En este capítulo introduciremos algunos principios importantes que rigen la evolución a nivel molecular y explicaremos cómo interpretar correctamente un árbol filogenético. A continuación, presentaremos los métodos de inferencia filogenética más frecuentemente utilizados, haciendo énfasis en los métodos probabilísticos por ser los que mejor extraen la información filogenética de las secuencias. Además, veremos cómo estimar la robustez estadística del árbol inferido y cómo el contraste de hipótesis puede utilizarse tanto para comparar árboles filogenéticos como para comprender cómo ha sido el proceso evolutivo subyacente.

### 9.1.1. Teoría de la Evolución

En 1859, veinte años después de regresar de su viaje a bordo del *Beagle*, Charles Darwin publicó *El origen de las especies*, donde formulaba detalladamente la teoría de la evolución. Este mérito debe ser compartido con Alfred Russell Wallace, que de forma independiente propuso la selección natural como el mecanismo generador de la diversidad biológica.

Darwin propuso que todos los organismos comparten un ancestro común, a partir del cual ha surgido toda la diversidad que observamos, y que los organismos similares se encuentran más emparentados entre sí, es decir, comparten un ancestro común más cercano. La clave del proceso evolutivo se encuentra en la selección natural. La teoría de la evolución propone que ciertos cambios espontáneos generan una variabilidad entre los organismos que puede heredarse. Las cualidades que resultan más favorables de cara a la adaptación al entorno (el factor determinante es el éxito reproductivo) serán las favorecidas por el mecanismo de selección natural. Hoy en día puede parecernos una idea sencilla, casi intuitiva, pero considerando que en aquel tiempo no se conocía el DNA ni las proteínas, es comprensible que se admire a Darwin y también que su teoría despertase tanta polémica.

El descubrimiento de los genes lo debemos a Gregor Mendel, cuyo trabajo hacia 1866 permitió conocer con mayor detalle cómo se produce la transmisión de información entre generaciones y cómo se genera la variabilidad sobre la que actuará la selección natural. Más tarde se determinó la naturaleza de estos genes como ácidos nucleicos (DNA) y de sus productos (proteínas), se descubrió la estructura de doble hélice y se descifró el código genético. Posteriormente fue posible determinar las secuencias de proteínas primero y DNA después, dando lugar a los primeros estudios de evolución molecular.

### 9.1.2. Evolución molecular

Gracias a los hallazgos anteriormente citados, hoy en día conocemos la base molecular de la evolución propuesta por Darwin. Las mutaciones del DNA son la fuerza productora de variabilidad sobre la que actúa la selección natural. La mutación se considera un proceso aleatorio, aunque está influenciado por una amplia variedad de factores, como lo son el modo de vida de los organismos, el tiempo de generación, o la tasa metabólica [2]. Además, genes distintos tienen tasas evolutivas distintas que dependen de la estructura y función de las proteínas que codifican. Distintas regiones del genoma presentan patrones y tasas de mutación diferentes, relacionados con el estrés replicativo (número veces que se copia por generación), la recombinación, la metilación del DNA o el posicionamiento de los nucleosomas. Debido a la naturaleza poblacional del proceso evolutivo, algunas de estas mutaciones se perderán con el tiempo, pero otras se fijarán en la población (lo que conocemos como sustituciones). En lo que respecta a la

inferencia filogenética, las sustituciones son una fuente de información muy valiosa que nos permite establecer relaciones de parentesco y estudiar la naturaleza del proceso evolutivo.

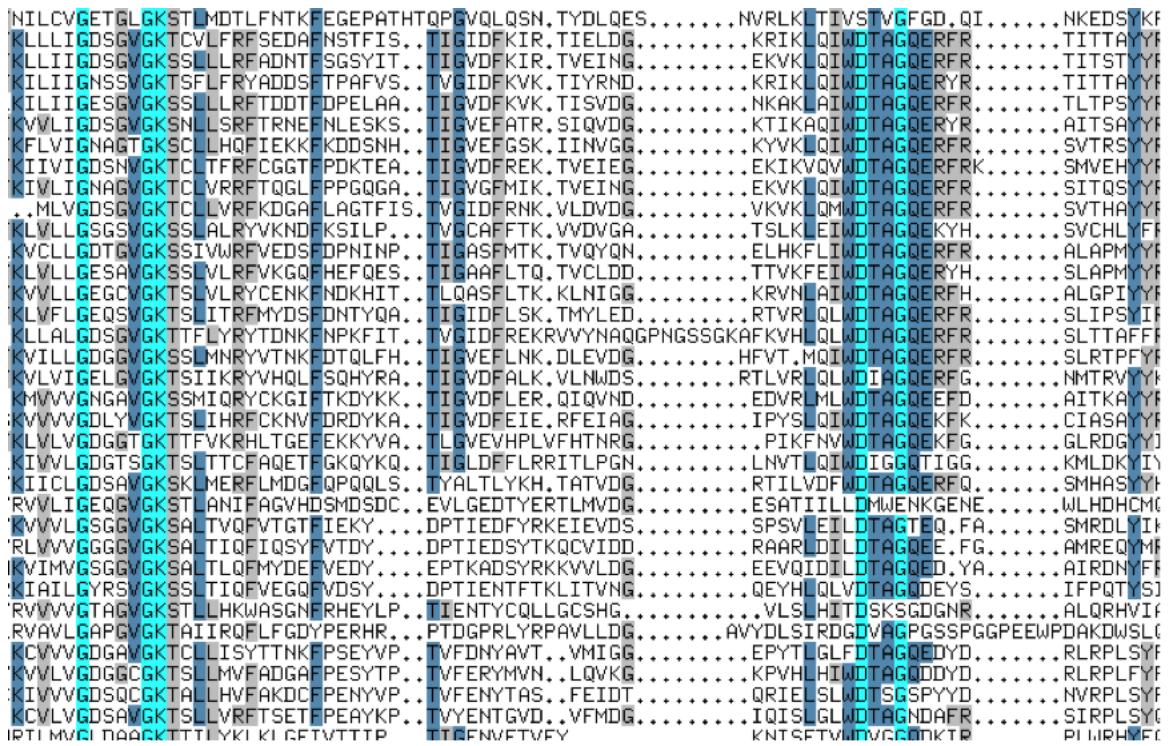
A lo largo de este capítulo nos centraremos en el análisis filogenético de secuencias de DNA o proteínas, puesto que presentan numerosas ventajas con respecto a otro tipo de datos y son (con diferencia) los más utilizados en la actualidad. Los primeros análisis comparativos de secuencias de proteínas, hace ya más de 50 años, supusieron un cambio importante en el enfoque del estudio de la evolución. En un principio se extrapolaron los principios evolutivos que se conocían del estudio de la morfología, aplicándolos al nivel molecular. A nivel morfológico todo cambio es objeto de *selección natural*, sea ésta positiva o negativa. La *selección positiva* (o evolución adaptativa) se refiere al proceso por el que los cambios que ofrecen cierta ventaja son favorecidos y tienden a fijarse en la población. Un ejemplo morfológico clásico es el de la evolución del cuello de la jirafa: un cuello más largo permite a los individuos que lo poseen comer las hojas de los árboles que otros no alcanzaban, aumentando sus posibilidades de dejar descendencia. La *selección negativa* (o purificadora) se refiere a la eliminación de aquellos cambios que perjudican la adaptación, y tienden a desaparecer de la población al reducir el éxito reproductivo.

Cuando Zuckerkandl y Pauling [48] compararon las primeras secuencias de globinas de distintas especies, observaron algo que no encajaba con esta visión: el número de cambios aminoácídicos entre linajes era aproximadamente proporcional al tiempo de divergencia estimado a partir del registro fósil. Este fenómeno se dio a conocer como reloj molecular. El reloj molecular parecía contradecir la visión existente de la evolución, según la cual todo cambio era objeto de selección. Como contraposición esta la visión seleccionista, Kimura (1983) [23] formuló la teoría neutral de la evolución molecular, según la cual una parte considerable de los cambios no afectan el éxito reproductor del individuo (o su efecto es muy pequeño), y al no estar sujetos a selección, estos cambios se acumulan a un ritmo aproximadamente constante a lo largo del tiempo. En la actualidad se asume que la evolución molecular está gobernada tanto por procesos neutrales como por la selección natural, si bien no siempre es sencillo conocer el peso específico de estas fuerzas evolutivas.

Los alineamientos múltiples son una forma de establecer relaciones de homología (origen evolutivo común) entre las posiciones (nucleótido o aminoácidos) de un conjunto de secuencias. La huella de algunos procesos evolutivos puede observarse fácilmente en los alineamientos múltiples. Así, la *selección negativa* se manifiesta como conservación en el alineamiento (todas las secuencias tienen el mismo nucleótido o amino ácido). Cuando una posición está conservada no es que no haya sufrido mutaciones, sino que cuando ha mutado la selección negativa (purificadora) ha actuado. La huella de la selección negativa es un buen indicio de relevancia funcional. La *evolución neutral*, en cambio, se manifiesta como variabilidad. Las posiciones que varían más son aquellas que tienen menor relevancia en el mantenimiento de la estructura y la función de la proteína. En realidad, entre estos dos extremos, conservación y variabilidad, existe todo un gradiente que depende de la relevancia de cada posición para la estructura y función del gen en cuestión (Figura 9.1). La huella de los procesos evolutivos se aprecia de forma diferente a nivel molecular y morfológico. A nivel molecular, por ejemplo, y a diferencia de lo que ocurre con la selección negativa y la evolución neutral, la huella de la *selección positiva* es difícil de detectar (volveremos a este problema más adelante). Por el contrario, a nivel morfológico la evolución positiva y negativa resultan obvias, mientras que la evolución neutral no se aprecia fácilmente.

### 9.1.3. Interpretación de un árbol

Un *árbol* consta de ramas y nodos que conectan estas ramas entre sí. Las *ramas terminales* (hojas) corresponden a las secuencias actuales (observadas), y las *ramas internas* representan sus ancestros hipotéticos. Su longitud refleja la cantidad de cambio acumulado. Los *nodos* relacionan estas ramas entre sí según su relación ancestro-descendiente. La rama más interna, el ancestro común, representa la

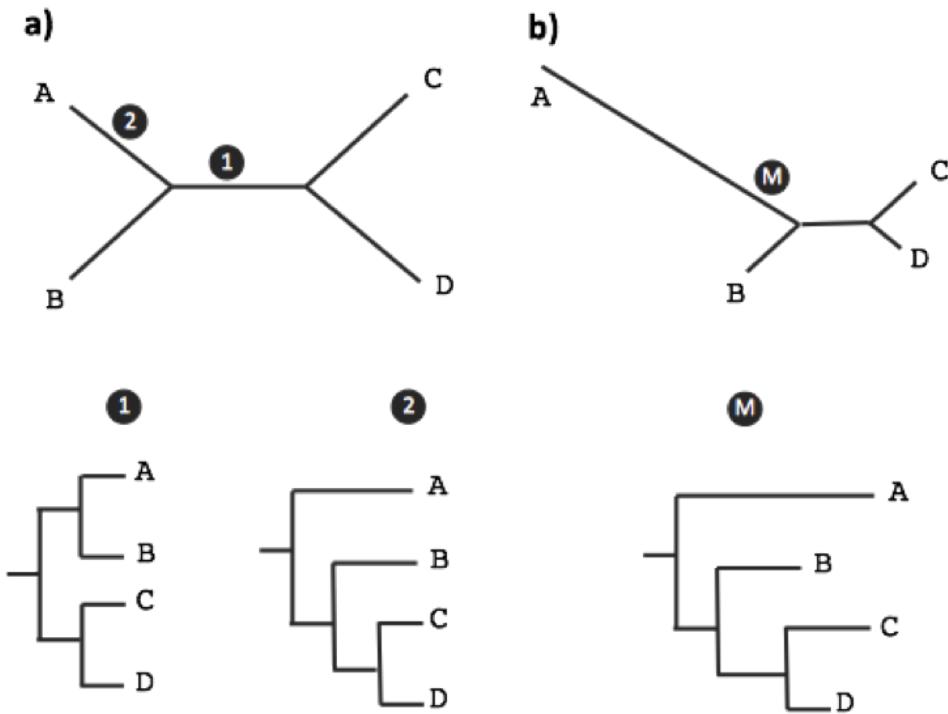


**Figura 9.1:** Alineamiento múltiple de proteínas. En él podemos observar regiones bajo distintas intensidades de selección que se manifiestan en posiciones más o menos conservadas en las distintas secuencias.

raíz del árbol y es esencial para dotar al árbol de un sentido histórico. Los métodos de reconstrucción filogenética a partir de secuencias (salvo alguna excepción) producen árboles sin raíz. En estos árboles, la raíz puede establecerse a posteriori en distintos puntos, lo cual determinará las relaciones de parentesco entre las secuencias. Por ejemplo, en la Figura 9.2, al enraizar el árbol en 1, la secuencia A será el grupo hermano de B, pero si lo enraizamos en 2, A será el grupo hermano de B, C, más D. Elegir un enraizamiento correcto es esencial para poder interpretar la historia evolutiva.

La forma más recomendable de enraizar un árbol es incluyendo una secuencia externa al grupo de interés que dote al árbol filogenético de sentido histórico (polaridad). La elección del *grupo externo* (o *outgroup*) requiere cierta información adicional, como la derivada de estudios morfológicos o taxonómicos. En algunos casos no disponemos de esta información o del grupo externo, en cuyo caso el árbol suele enraizarse en el *punto medio* (*midpoint rooting*), es decir, en la rama más larga (M en la Figura 9.2(b)). Cuando las secuencias evolucionan a velocidades muy distintas (ausencia de reloj molecular, ver más adelante) el enraizamiento en el punto medio debe interpretarse con precaución.

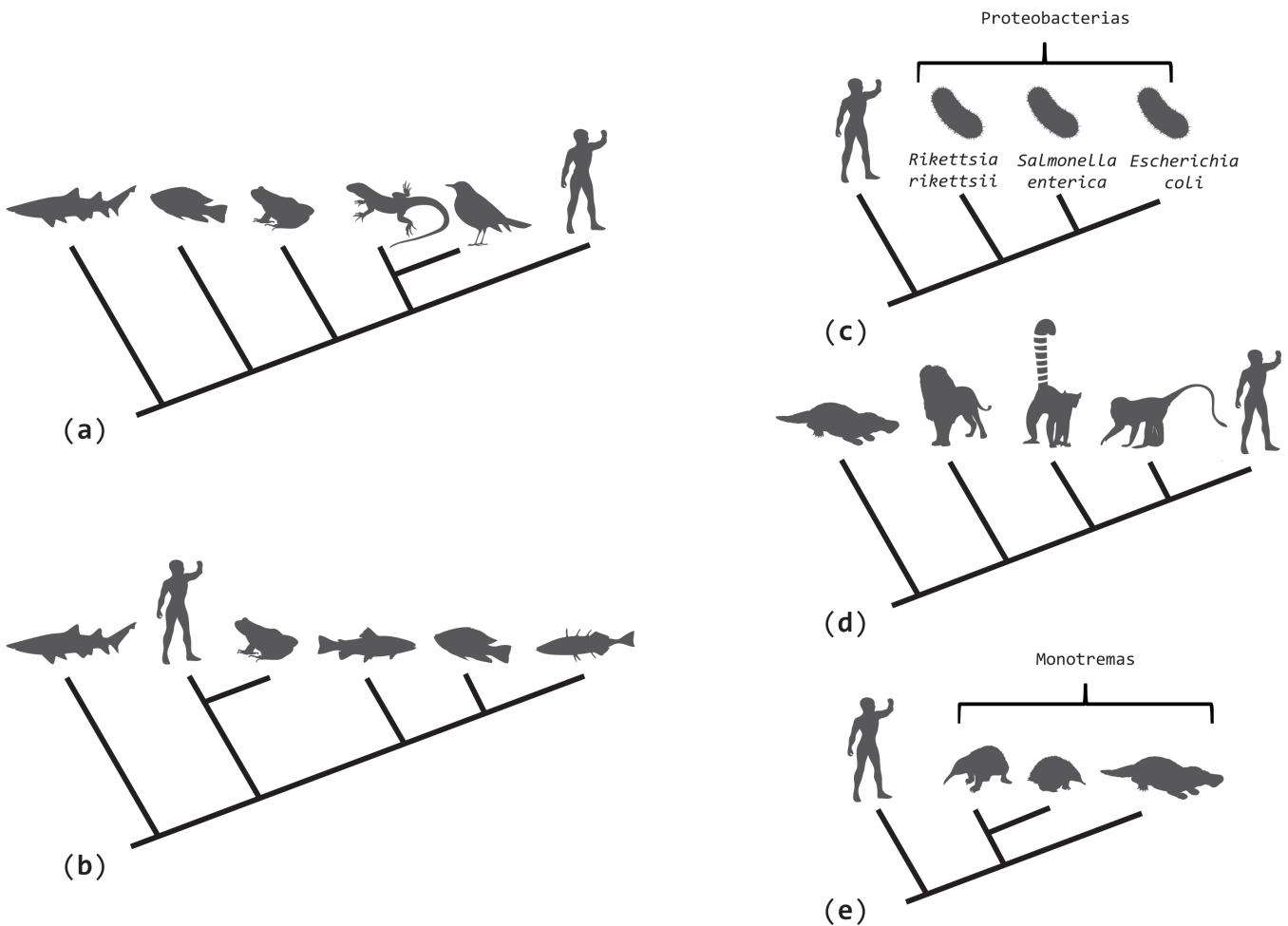
A pesar de que los árboles filogenéticos son representaciones bastante intuitivas de la historia evolutiva, ciertos errores de interpretación son relativamente frecuentes, incluso entre científicos experimentados [30]. El error más común es interpretar el árbol como una línea de progreso. Así, si observamos la Figura 9.3(a), podríamos caer en la tentación (antropocéntrica) de ver una direccionalidad en la evolución desde peces hasta humanos, pasando por anfibios, reptiles y aves. Sin embargo, una representación alternativa pero coherente con la anterior podría sugerirnos lo contrario (Figura 9.3(b)). Si 9.3(a) y 9.3(b) te sugieren relaciones diferentes es que estás interpretando erróneamente los árboles; posiblemente estás asumiendo líneas de progreso entre distintas especies actuales. Es un error pensar que entre las especies actuales (observables) hay unas más evolucionadas que otras; todas lo están por igual, simplemente representan distintas ramas del árbol de la vida. La línea de progreso existe, pero ésta



**Figura 9.2:** La raíz da sentido histórico (polariza) a un árbol filogenético. En (a) se muestran dos posibles enraizamientos (1, 2) para este árbol de cuatro secuencias (A, B, C, D) y cómo éstos se traducen en historias evolutivas diferentes. En (b) se muestra el enraizamiento en el punto medio (*midpoint rooting*).

en el tiempo: va desde la raíz del árbol hacia las hojas y no entre distintas hojas. Para evitar caer en este error, recuerda que las ramas de los árboles pueden rotarse de cualquier manera sin que ello altere las relaciones representadas en el árbol. Otro ejemplo clarificador lo encontramos en la Figura 9.3(c), donde el humano es el grupo externo utilizado para enraizar un árbol de bacterias. Ese árbol, aunque pueda sorprenderte, es absolutamente correcto; si te sorprende es porque asumes (consciente o inconscientemente) una direccionalidad evolutiva en el sentido equivocado. Este árbol simplemente nos dice que las bacterias están más cercanamente emparentadas entre sí de lo que lo están con los humanos, pero en ningún caso que los humanos hayan dado origen a las bacterias.

Hablando de errores comunes, es frecuente oír hablar de especies basales, primitivas o poco evolucionadas. La Figura 9.3(d) muestra un árbol de mamíferos placentarios donde el ornitorrinco (monotremo) es el grupo externo. Si al ver este árbol piensas que el ornitorrinco es un mamífero basal o primitivo, tal vez ver el árbol 9.3(e) te ayude a reconocer tu error. Para hablar con propiedad, debemos referirnos a grupos de divergencia temprana o tardía (*early or late branching*) [7] y referirnos a las relaciones representadas en el árbol como relaciones de grupos hermanos (*sister group relationships*).



**Figura 9.3:** Árboles filogenéticos que tratan de ilustrar ciertos errores comunes en su interpretación (ver texto).

## 9.2. Métodos de reconstrucción filogenética

En esta sección describiremos algunos de los métodos más utilizados para reconstruir la historia evolutiva a partir de datos de secuencia de DNA o proteínas, aunque estos métodos también pueden utilizarse con otro tipos de datos, como por ejemplo caracteres morfológicos.

Existen dos tipos principales de métodos de reconstrucción filogenética: (i) los que utilizan distancias genéticas, siendo los más conocidos el método *neighbor-joining* (NJ) y el de *mínima evolución* (ME); y (ii) los basados en caracteres, como los métodos de *máxima parsimonia* (MP), de *máxima verosimilitud* (MV) (en inglés, *maximum likelihood*) y de *inferencia Bayesiana* (IB) [18, 43]. Dentro de los métodos basados en caracteres, MV e IB (a diferencia de MP) son métodos probabilísticos que tratan la inferencia filogenética como un problema estadístico, y utilizan modelos explícitos de evolución molecular para el cálculo de probabilidades. Además de los anteriores, existen otros métodos, pero debido a la gran variedad y complejidad de éstos quedan fuera del alcance de este capítulo. Aquí introduciremos únicamente los métodos más comúnmente utilizados: un método que utiliza distancias genéticas (NJ) y los métodos probabilísticos (MV e IB). Los métodos probabilísticos son los que mejor aprovechan la información filogenética contenida en las secuencias, los más avanzados y generalmente los más fiables.

Para saber más acerca de métodos filogenéticos: Swofford et al. 1996 [43], Felsenstein 2004 [14] y Lemey et al. 2009 [27].

### 9.2.1. El alineamiento múltiple, la información de partida

Los *alineamientos múltiples* representan hipótesis de homología de posición entre caracteres (nucleótidos, codones o aminoácidos) de las distintas secuencias. Son el punto de partida para inferir filogenias y estudiar patrones y tasas de cambio evolutivo. Existen muchos métodos para el alineamiento de secuencias y han sido tratados en detalle en el Capítulo 8 “Análisis de secuencias biológicas”. Debemos tener en cuenta que partir de un buen alineamiento es clave para obtener estimas fiables de la filogenia. Pero *¿cómo saber si nuestro alineamiento es correcto?* Una cuestión particularmente importante es que además de la información filogenética, los alineamientos también contienen cierto ruido, por lo que es recomendable analizarlos cuidadosamente.

En ocasiones, como cuando alineamos regiones que evolucionan rápidamente, establecer la homología de posición resulta complejo. En las regiones más variables es común que hayan ocurrido múltiples cambios superpuestos (mutaciones sobre mutaciones), inserciones y delecciones. Estos fenómenos producen hipótesis de homología entre caracteres erróneas y el alineamiento no es fiable. Esto es el *ruido* al que nos referíamos, que puede incluso llegar a superar a la *señal filogenética*. Para aliviar estos problemas existen diversos métodos para *limpiar* alineamientos, los cuales básicamente eliminan las regiones más variables. GBlocks [5] retiene los bloques del alineamiento que contengan pocos gaps (inserciones y delecciones) y presenten poca variabilidad, pudiendo ajustarse los parámetros según el caso. TrimAl [4] es un método muy versátil que acepta múltiples opciones, como la eliminación de columnas del alineamiento con un porcentaje determinado de gaps o una similitud menor de 0,5 (Tabla 9.1). Estos métodos nos permiten filtrar los alineamientos y trabajar solo con aquellas hipótesis de homología entre caracteres más robustas, lo cual tiene un impacto muy positivo sobre cualquier análisis posterior, incluyendo la reconstrucción filogenética [5].

### 9.2.2. Modelos de evolución, el engranaje

Los *modelos evolutivos* son descripciones matemáticas de la evolución de las secuencias y constituyen el engranaje que nos permite conectar los datos (alineamientos) con los métodos de reconstrucción filogenética. En función de lo que consideremos la unidad de evolución, existen modelos para nucleótidos, aminoácidos o codones.

Generalmente asumimos que las posiciones de un alineamiento (nucleótidos, aminoácidos o codones) evolucionan de forma independiente, y aunque sabemos que ésta no es una asunción completamente realista, simplifica enormemente la complejidad computacional de inferir filogenias. La sustitución entre nucleótidos, aminoácidos o codones se modela como un evento aleatorio y los cambios entre los estados se describen mediante una cadena de Markov tiempo-continua [3]. La propiedad más importante de las cadenas de Markov es que no tienen memoria, es decir, que la probabilidad de cambio entre nucleótidos sólo depende del nucleótido actual y no de aquéllos nucleótidos que pudo haber en esa posición con anterioridad. Lo mismo se aplica también a los modelos de aminoácidos y codones. Además de estas asunciones, los modelos utilizados con mayor frecuencia (y que tratamos en este capítulo), asumen ciertas propiedades del proceso de evolución molecular, que se considera globalmente estacionario, reversible, y homogéneo (para mayor detalle: [20]).

Algunos de estos modelos son bastante sencillos (emplean pocos parámetros), mientras que otros son muy complejos. Los modelos tienen dos tipos parámetros principales: las probabilidades de cambio

entre estados (ya sean nucleótidos, aminoácidos, o codones) y las frecuencias de estos nucleótidos, aminoácidos, o codones en nuestro alineamiento. En los modelos de nucleótidos y codones, los valores de estos parámetros se estiman a partir de los datos observados.

El modelo de evolución de nucleótidos más sencillo es el conocido como *Jukes-Cantor* (JC) y asume que cualquier cambio entre nucleótidos tiene la misma probabilidad de ocurrir y que las frecuencias de los cuatro nucleótidos son iguales [21]. Bajo las asunciones de estacionariedad, reversibilidad y homogeneidad, el modelo más complejo es GTR (del inglés *general time reversible*), que utiliza un parámetro distinto para modelar cada una de las sustituciones entre nucleótidos y sus frecuencias [44]. Entre ambos extremos tenemos muchos otros modelos, como K80, F81, F84, HKY85 o TN93 [47].

En el caso de los modelos de reemplazamiento de aminoácidos, en cambio, los parámetros generalmente no se estiman a partir de los datos, sino que vienen previamente fijados. Decimos que son *modelos no paramétricos*. La razón para ello es sencilla: al ser el alfabeto de aminoácidos más extenso que el de nucleótidos (20 vs 4), el número de parámetros a estimar se incrementa mucho y su estimación a partir del alineamiento produciría estimas con varianzas elevadas, lo cual no es conveniente. Por esta razón se utilizan matrices de reemplazamiento precalculadas, que han sido estimadas a partir de grandes conjuntos de datos [47]. Algunas de estas matrices son Blosum, Dayhoff, JTT, LG, MtArt, MtREV o WAG. En algunos casos, se permite utilizar las frecuencias observadas de aminoácidos, lo que añade 19 parámetros al modelo (se indica como +F en la selección de modelos; ver más abajo).

Además de modelar las probabilidades de cambio entre nucleótidos (o aminoácidos) y sus frecuencias relativas, con frecuencia se utilizan parámetros adicionales para tener en cuenta el hecho de que distintas partes del alineamiento pueden evolucionar a tasas distintas. En muchos casos, la adición de estos parámetros incrementa notablemente la capacidad del modelo para explicar nuestros datos, mejorando así la inferencia filogenética [47]. La variación de tasas entre sitios puede tenerse en cuenta de formas diferentes. La primera es añadir un parámetro para modelar la variación de tasas entre las posiciones del alineamiento. Esta variación se describe mediante una distribución gamma que se especifica con un único parámetro (alfa) [47]. Este parámetro podemos encontrarlo en la forma +G, “*gamma shape*”, “*alpha shape*” o “*rate variation across sites*”. La segunda forma es asumir que una proporción de las posiciones en nuestro alineamiento son invariables (completamente conservadas) [35]. Este parámetro podemos encontrarlo como +I, “*p-inv*” o “*proportion of invariable sites*”. Ambos parámetros pueden también utilizarse conjuntamente (+I+G).

Ante la gran variedad de modelos posibles, la pregunta obvia es *¿qué modelo debemos elegir?* La respuesta es que depende del conjunto de datos que estemos analizando. Para cada conjunto de datos es necesario identificar el modelo que mejor se ajuste, pues cuanto más cercano sea éste a la realidad de los datos, más posibilidades tendremos de inferir una buena filogenia. Cuantos más factores queramos tener en cuenta, más parámetros tendremos que incluir en el modelo. En principio, uno podría estar tentado de elegir siempre el modelo más complejo, pero entonces nos topamos con un problema estadístico. A mayor número de parámetros, mayor será la varianza con la que se estimen, es decir, las estimas serán menos fiables. Por el contrario, un modelo sencillo (con pocos parámetros) permitirá hacer estimas muy estables, pero se ajustará peor a los datos. Por tanto, es necesario encontrar un balance entre el ajuste a nuestros datos y la complejidad del modelo.

Existen distintos criterios que pueden utilizarse para elegir el modelo que mejor se ajusta a nuestros datos, al mismo tiempo que sopesamos su complejidad (el número de parámetros). Tradicionalmente se han venido utilizando los contrastes basados en la razón de verosimilitudes (*likelihood ratio test*, LRT). Los LRT comparan modelos por pares y requieren que las hipótesis estén anidadas, es decir, que el modelo nulo (aquel con menos parámetros) sea un caso particular del modelo alternativo [19]. Estos contrastes son muy útiles para testar ciertos tipos de hipótesis filogenéticas (ver la Sección 9.3), aunque

se ha desaconsejado su uso para la selección de modelos evolutivos en favor de otras alternativas como el criterio de información de Akaike (*Akaike information criterion*, AIC) [34]. El AIC es una medida de la cantidad de la información que perdemos al explicar los datos observados utilizando un modelo en particular. Por tanto, el modelo que mejor se ajusta a los datos observados es aquél que tiene un valor de AIC menor. A diferencia del LRT, el AIC permite comparar más de dos modelos simultáneamente y no requiere que estén anidados. El AIC se calcula como:

$$AIC = 2k - 2 \ln L \quad (9.1)$$

donde  $L$  es el valor de verosimilitud y  $k$  el número de parámetros del modelo.

La selección del modelos en función de AIC (y otros criterios) puede hacerse mediante los programas jModelTest (para secuencias de nucleótidos) [32], ProtTest (para secuencias de aminoácidos) [1] o ModelGenerator (para ambos) [22]. Estos programas calculan la verosimilitud de todos los modelos para un alineamiento dado y utilizan el AIC (u otro criterio) para identificar el modelo que mejor se ajusta a los datos. Para mayor detalle acerca de la selección de modelos leer el texto de David Posada [33].

En la actualidad es frecuente utilizar conjuntos de datos que combinan varios genes, con el objeto de tener grupos de datos independientes y más caracteres totales con los que reconstruir la historia evolutiva. En estos casos, en lugar de utilizar un único modelo para el conjunto de datos global, se suelen utilizar modelos evolutivos independientes para cada gen, de modo que se puedan tener en cuenta sus distintas propiedades evolutivas. Este tipo de *particiones* se puede aplicar no sólo a genes, sino a regiones de éstos (por ejemplo hélices transmembrana) o a posiciones de los codón (por ejemplo usando un modelo diferente para las tercera posiciones, que evolucionan a mayores tasas). Frecuentemente, el número de particiones posibles es enorme, por lo que es preciso utilizar un criterio explícito para su elección. Al igual que en la selección de modelos evolutivos, podemos utilizar el valor de AIC para elegir el esquema de partición que mejor se ajusta a nuestros datos, para lo cual podemos utilizar el programa PartitionFinder [26].

Todos los modelos de evolución molecular realizan ciertas asunciones, pero gracias a que éstas son explícitas, pueden ser evaluadas estadísticamente (trabajamos en un marco probabilístico) y así podemos comprender mejor el proceso evolutivo. Por eso no sólo existen modelos para optimizar la reconstrucción filogenética, sino también otros que nos permiten caracterizar el proceso evolutivo en sí mismo. Así, estos otros modelos, que veremos en la Sección 9.3, nos permiten responder a preguntas tales como si ha habido selección positiva o si dos caracteres han coevolucionado.

### 9.2.3. Métodos de distancias: rápidos pero no tan fiables

La forma más sencilla de medir la distancia genética es contar el número de diferencias entre dos secuencias, lo que conocemos como *distancia observada* (o distancia  $p$ ), que se expresa como el número de diferencias por posición. Pero esta forma de calcular distancias tiene un problema: a medida que aumenta la divergencia entre las secuencias, la probabilidad de que ocurran cambios superpuestos es también mayor. Los *cambios superpuestos* enmascaran cambios previos ocurridos en esa posición, por lo que la distancia observada en estos casos es una subestimación de la distancia real. Además, por azar, puede ocurrir que cambios superpuestos hayan dado lugar al mismo nucleótido en dos secuencias distintas. En ese caso, la observación de un mismo nucleótido no se debería a descendencia común, sino que representaría una convergencia, dando lugar a lo que conocemos como *homoplásia*. La homoplásia introduce ruido en los datos, afectando seriamente a la estimación de las distancias genéticas y la

inferencia filogenética. A medida que aumenta la divergencia entre dos secuencias, aumenta también la subestimación del número real de cambios calculados como distancias observadas. Cuando esto ocurre, decimos que las secuencias están saturadas o que hay *saturación*. Los métodos basados en distancias pueden utilizar los modelos de evolución para intentar corregir las distancias observadas y aliviar el problema de la saturación.

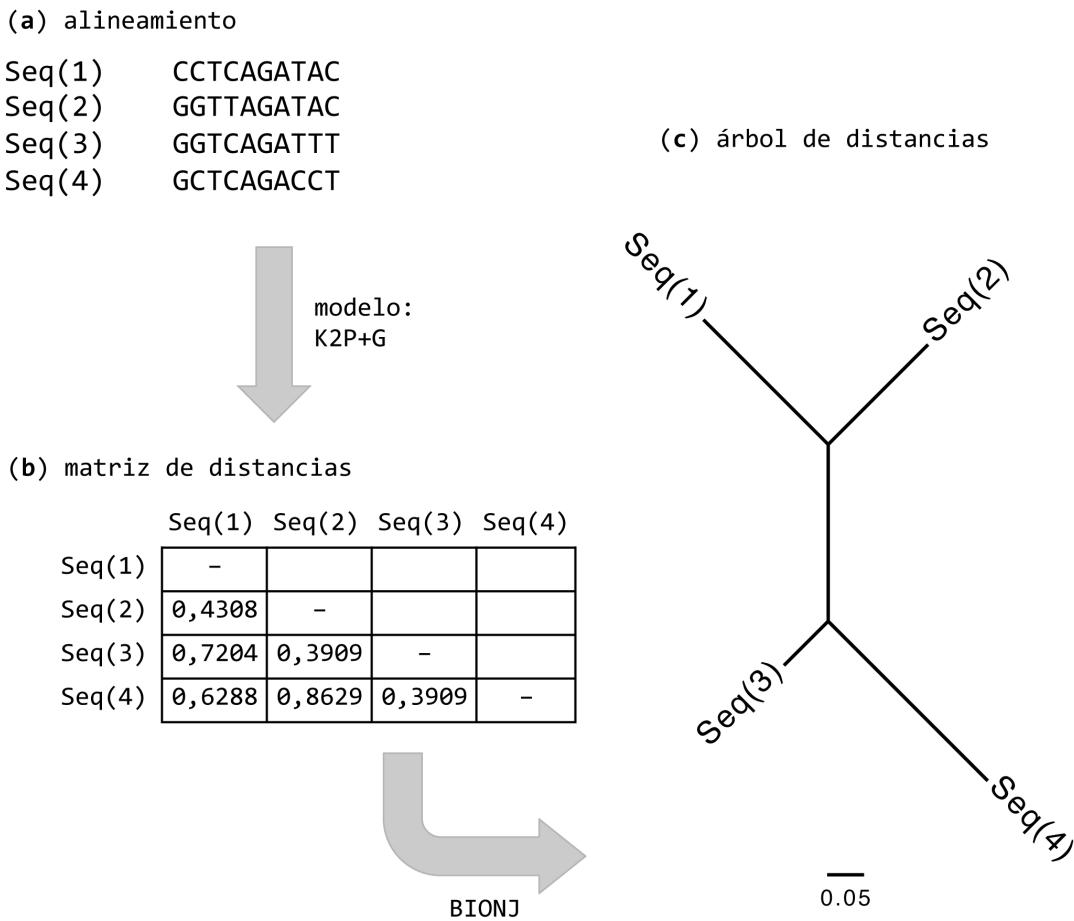
Los métodos que reconstruyen árboles filogenéticos a partir de distancias genéticas operan en dos pasos (Figura 9.4). En un primer lugar, se calculan las distancias genéticas entre todos los pares de secuencias, y se resume esta información en una matriz de distancias. Posteriormente, se utilizan los valores de esta matriz para reconstruir un árbol filogenético. Idealmente, las distancias genéticas han de reflejar las divergencias evolutivas reales.

Si asumiéramos que todas las secuencias han evolucionado con la misma tasa (lo que conocemos como la *hipótesis del reloj molecular*), las distancias genéticas serían directamente proporcionales al tiempo. Asumiendo además que las distancias genéticas reflejan divergencia real entre las secuencias, la reconstrucción filogenética se convertiría en un problema trivial (una simple regla de tres) y podríamos resolverlo con un método sencillo. Sin embargo, las distancias genéticas raramente son un buen estimador de la divergencia real y en muchos casos, la tasa de cambio en distintas secuencias es variable.

La reconstrucción de un árbol filogenético a partir de una matriz de distancias puede hacerse utilizando varios algoritmos. Uno de los métodos de distancias más conocido es *Neighbour-Joining* (NJ) [38] y una de sus variantes es BIONJ [15], que tiene en cuenta el hecho de que divergencias mayores acarrean varianzas mayores. Estos métodos no asumen la existencia de un reloj molecular para la reconstrucción filogenética.

Los métodos de distancias tienen la ventaja de ser muy rápidos, y resultan útiles como parte de un análisis preliminar. Los árboles estimados con métodos de distancias pueden ser similares o iguales al árbol estimado con métodos probabilísticos si partimos de unos datos limpios (con poco ruido), pero tienen un rendimiento bastante malo cuando las asunciones del modelo evolutivo son violadas por los datos [43]. Además, los métodos de distancias no estiman la confianza que tenemos sobre el árbol generado, es decir, no nos dicen si existen hipótesis alternativas (otros árboles) que pudieran explicar las relaciones entre las secuencias de modo similar. El mayor problema de los métodos de distancias es que pierden gran cantidad de información al resumir todas las diferencias entre las secuencias del alineamiento en un único número (la distancia genética), y también que son más sensibles al problema de atracción de ramas largas (ver más adelante). Una vez calculada la distancia genética, los métodos de distancias no nos permiten ir hacia atrás para ver en qué parte del árbol ocurrió un determinado cambio en las secuencias.

Los métodos como NJ producen siempre un único árbol que viene determinado por la matriz de distancias. Debido a los problemas asociados a las distancias genéticas que hemos mencionado, los métodos como NJ pueden generar árboles que difieran de la filogenia verdadera. Además, el algoritmo de NJ no garantiza encontrar el mejor árbol (según la matriz de distancias), pudiendo quedar atrapado en mínimos locales. En estos casos, es posible encontrar árboles similares mediante pequeñas reorganizaciones locales (conocidas como búsquedas heurísticas; ver más adelante) y utilizar un criterio de optimización para decidir qué árbol, dentro del conjunto de árboles similares, es el más adecuado. El método de mínima evolución, por ejemplo, prefiere el árbol más corto (aquél con la suma de todas longitudes de rama menor) como explicación de la historia evolutiva [37].



**Figura 9.4:** Reconstrucción filogenética a partir de distancias genéticas. Partiendo de un alineamiento múltiple (a), se calcula una matriz de distancias (b) que es utilizada para estimar el árbol filogenético (c), en este caso mediante BIONJ.

#### 9.2.4. Métodos probabilísticos basados en caracteres

A diferencia de los métodos basados en distancias, los basados en caracteres tienen en cuenta los cambios que ocurren en cada posición del alineamiento, y hacen por tanto un uso más eficiente de la información [43]. Su funcionamiento se fundamenta en *criterios de optimización*: en un primer paso se generan árboles utilizando ciertos algoritmos heurísticos, y en un segundo paso estos árboles se evalúan utilizando una función objetiva que nos permite elegir el mejor árbol entre todos los generados en el paso anterior.

En principio, lo deseable sería poder comparar todos los árboles posibles utilizando nuestro criterio de optimización e identificar cuál de entre todos ellos es el mejor. Sin embargo, evaluar todos los posibles árboles es generalmente muy poco práctico, ya que este número crece de forma factorial con respecto al número de secuencias. Por ejemplo, existen tres árboles sin raíz posibles para explicar las relaciones entre cuatro secuencias, pero el número de árboles posibles para 10 secuencias es de más de dos millones, y para 80 secuencias aumenta hasta  $2,18 \times 10^{137}$  (un número mucho mayor que el número estimado de protones en el Universo,  $10^{89}$ !). En la práctica, no es factible comparar todos los árboles posibles si

no es para un número de secuencias menor de 6-7. De forma general, se utilizan *métodos heurísticos* que aunque no garantizan encontrar el mejor árbol globalmente, pero permiten evaluar un número significativo de árboles más probables. Las búsquedas heurísticas utilizan algoritmos que partiendo de un árbol (por ejemplo, aleatorio o reconstruido con NJ) realizan pequeñas reorganizaciones locales para encontrar árboles más probables. Durante las búsquedas heurísticas, las distintas reorganizaciones se aceptan o rechazan con ayuda del criterio de optimización, y la búsqueda continúa hasta que no se encuentren más mejoras significativas.

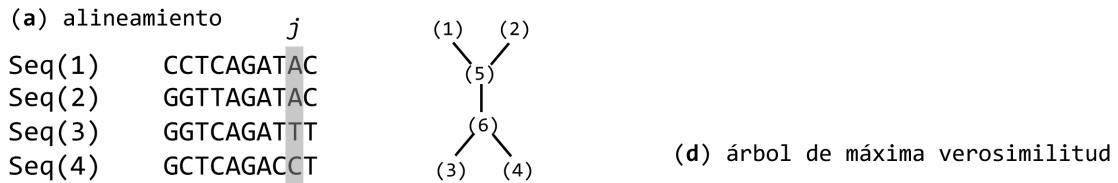
## Máxima verosimilitud

La aplicación del principio de *máxima verosimilitud* (MV) a la reconstrucción filogenética [10, 11] supuso un gran avance al permitir estudiar las relaciones evolutivas desde un punto de vista estadístico. Al trabajar en un *marco probabilístico*, es posible estimar patrones históricos e inferir parámetros intrínsecos a los procesos de evolución molecular, y el contraste de hipótesis filogenéticas se realiza de forma directa. El principio de MV se define como la probabilidad de que un modelo de evolución dado (de nucleótidos, codones o aminoácidos) y una historia evolutiva hipotética (el árbol) hayan dado lugar a los datos observados (alineamiento de secuencias) [11]. El árbol de MV es aquél que tiene la mayor probabilidad de haber generado los datos observados, y por tanto se prefiere como hipótesis filogenética. El método de MV es eficiente y consistente. Esto significa que la varianza de las estimas es menor que en otros métodos (están menos afectadas por el error de muestreo) y que las estimas convergen al valor correcto del parámetro conforme aumentamos el número de caracteres [14]. Estudios de simulación han demostrado que la reconstrucción filogenética por MV es bastante robusta, aunque no inmune, a posibles violaciones de las asunciones del modelo por parte de los datos [43].

La MV trata de inferir la *historia evolutiva más probable*. Como hemos visto anteriormente, para calcular esta probabilidad es necesario utilizar los *modelos evolutivos*, pues son el engranaje que describe la evolución de las secuencias en términos de probabilidad. A continuación veremos el fundamento del cálculo del valor de verosimilitud de forma intuitiva y su aplicación a la reconstrucción filogenética. Los modelos evolutivos generalmente asumen que el proceso de evolución molecular es reversible. Esto permite que el valor de verosimilitud de un árbol sea independiente de la posición de la raíz [40]. Además, hemos asumido que las posiciones evolucionan de forma independiente en el alineamiento. Esta asunción nos permite calcular la verosimilitud de forma independiente para cada posición del alineamiento y combinar posteriormente estas verosimilitudes en un valor final.

El cálculo de verosimilitud para cada posición se realiza teniendo en cuenta todos los posibles escenarios que han podido dar lugar a los datos observados (para esa posición) [43]. Esto implica sumar las probabilidades de cada combinación única de estados ancestrales (reconstrucciones hipotéticas en los nodos internos) y longitudes de rama (Figura 9.5). Obviamente, ciertos escenarios son mucho más probables que otros. Por ejemplo, en la Figura 9.5(b), la probabilidad de encontrar A en el ancestro de 1 y 2 (nodo 5) es mayor que la probabilidad de encontrar C, G, o T. Sin embargo, al calcular la verosimilitud se contemplan todas las posibilidades aunque su probabilidad sea muy baja. Puesto que la probabilidad de un escenario concreto (entre todos los posibles) es extremadamente pequeña, y que éstos valores de verosimilitud son a su vez el producto de muchas probabilidades, en la práctica trabajamos con los logaritmos naturales de las verosimilitudes ( $\ln L$ ). Una vez que hemos calculado la verosimilitud para cada posición en el alineamiento, la verosimilitud total del árbol se calcula como la suma de los logaritmos de las verosimilitudes para cada sitio (Figura 9.5).

La verosimilitud no debe confundirse con una probabilidad, a pesar de que se define en términos probabilísticos. La *verosimilitud* es la probabilidad del evento observado, que se calcula utilizando un modelo evolutivo que asumimos como correcto. Sin embargo, la verosimilitud no tiene nada que ver con



(d) árbol de máxima verosimilitud

(b) verosimilitud para el sitio  $j$

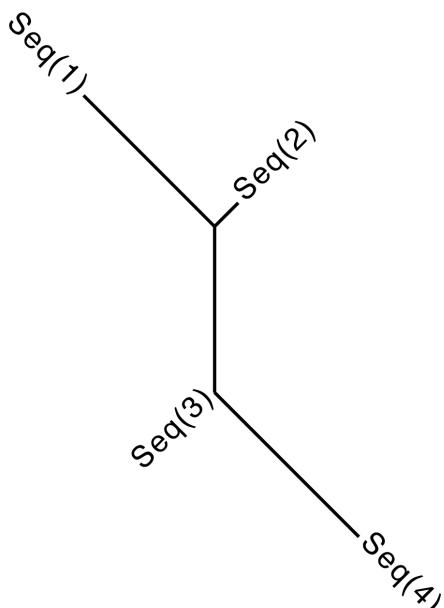
$$L(j) = \text{prob} \begin{bmatrix} A & A \\ A & \\ \diagdown & \diagup \\ T & C \end{bmatrix} + \text{prob} \begin{bmatrix} A & A \\ A & \\ \diagdown & \diagup \\ T & T \end{bmatrix} + \text{prob} \begin{bmatrix} A & A \\ A & \\ \diagdown & \diagup \\ T & C \end{bmatrix} +$$

$$+ \text{prob} \begin{bmatrix} A & A \\ A & \\ \diagdown & \diagup \\ T & G \end{bmatrix} + \text{prob} \begin{bmatrix} A & A \\ A & \\ \diagdown & \diagup \\ T & T \end{bmatrix} + \dots + \dots$$

(c) verosimilitud para el alineamiento

$$\ln L(\text{total}) = \ln L(1) + \ln L(2) + \dots + \ln L(N) = \sum_{j=1}^N \ln L(j)$$

0.08



**Figura 9.5:** Reconstrucción filogenética por máxima verosimilitud. Dados unos datos y un árbol (a), se calculan los valores de verosimilitud para cada sitio de forma independiente (b), y éstos se combinan en un valor final (c). El valor de verosimilitud nos permite evaluar los árboles generados durante la búsqueda heurística y elegir el árbol de máxima verosimilitud (d).

que el modelo que asumimos describa correctamente o no el proceso evolutivo real. De ahí la importancia de elegir el modelo que mejor se ajusta a nuestros datos. Es más, el hecho de que obtengamos distintos valores de verosimilitud al asumir modelos distintos es el fundamento de la selección de modelos y el contraste de hipótesis basado en verosimilitud ([19, 33]; volveremos a ello más adelante).

Como decíamos antes, los métodos de búsqueda heurística nos permiten “visitar” un cierto número de árboles entre todos los posibles, y valorar si un árbol es mejor o peor en función de un criterio de optimización, en este caso su verosimilitud. El árbol de MV será aquél que tiene el mayor valor de verosimilitud (menor  $\ln L$ ) entre todos los árboles evaluados durante la búsqueda heurística. Esperamos que la búsqueda heurística nos permita encontrar el mejor árbol entre todos los posibles (alcanzar el máximo global). Sin embargo, una búsqueda heurística nunca nos asegurará que hemos encontrado el mejor árbol, pues durante el proceso de optimización puede quedar atrapada en un máximo local, lo cual nos generará un árbol subóptimo. Para evitar este problema, podemos realizar varias búsquedas de MV (por ejemplo, 10 o 100) a partir de otros tantos árboles iniciales. Realizar múltiples búsquedas heurísticas nos permite explorar de forma más eficiente el conjunto de árboles posibles, y en última instancia, nos permite comparar los distintos árboles de MV encontrados en cada búsqueda heurística.

Si las diferentes búsquedas heurísticas (que parten de puntos independientes) coinciden en los árboles de MV obtenidos, tendremos mayor certeza de que hemos explorado bien el conjunto de posibles árboles, aumentando por tanto la confianza en la robustez de nuestra estima.

Existen multitud de programas con los que podemos realizar análisis de MV. Entre los más conocidos y utilizados están PhyML o Tree-puzzle, y otros especialmente útiles cuando utilizamos alineamientos grandes, como son RAxML o GARLI (Tabla 9.1).

## Métodos Bayesianos

La estadística Bayesiana, a diferencia de la estadística clásica, considera los parámetros de un modelo como variables aleatorias con una distribución estadística y no como constantes desconocidas. Puesto que desconocemos los valores de los parámetros del modelo, desde un punto de vista Bayesiano resulta más razonable especificar a priori unas distribuciones para describir los valores que éstos puede tomar. La *inferencia Bayesiana* (IB) de la filogenia se basa en el cálculo de la probabilidad posterior de un árbol, esto es, la probabilidad de que dicho árbol sea correcto dados unos datos y un modelo. La probabilidad posterior se calcula mediante el *teorema de Bayes*, según el cual la probabilidad posterior de un árbol  $P(T, \theta|D)$  es proporcional a su probabilidad previa  $P(T, \theta)$  multiplicada por su verosimilitud  $P(D|T, \theta)$  (o probabilidad de los datos dados un árbol y modelo evolutivo):

$$P(T, \theta|D) = \frac{P(T, \theta)P(D|T, \theta)}{P(D)} \quad (9.2)$$

donde el numerador  $P(D)$ , que corresponde a la probabilidad marginal de los datos, que es un factor de normalización. La probabilidad marginal de los datos es una suma sobre todas las posibles topologías, y para cada topología una integral sobre todas las longitudes de rama y parámetros del modelo, cuya función es que la suma de probabilidades posteriores de los árboles y sus parámetros sea uno.

Para entender mejor cómo funciona la IB, observemos el gráfico en la Figura 9.6(b), donde se representan en una única dimensión un conjunto de árboles posibles (eje x) en función de su probabilidad posterior (eje y). En la curva, cada punto representa una combinación única de topología y longitudes de rama. Obviamente, esta representación no es real, ya que el conjunto de árboles posibles es mucho más complejo y la distribución de las probabilidades posteriores es multimodal. La IB trata de determinar el área bajo esta curva, que corresponde a la probabilidad posterior de cada árbol concreto. En la Figura 9.6(b), podemos observar tres regiones de probabilidad posterior elevada, que corresponden a tres topologías distintas con su combinación particular de longitudes de rama.

Como hemos dicho anteriormente, el cálculo de la probabilidad posterior implica evaluar todos los posibles árboles y, para cada árbol, investigar todas las posibles combinaciones de longitudes de ramas y parámetros del modelo evolutivo. De nuevo, este cálculo resulta prohibitivo computacionalmente. En el caso de la IB, utilizamos algoritmos MCMC (o *Markov Chain Monte Carlo*) para obtener una aproximación de la distribución de la probabilidad posterior [16, 28]. Los MCMC son algoritmos de simulación estocástica que evitan el cálculo directo de las probabilidades posteriores y permiten obtener una muestra de la distribución posterior. De manera análoga a cómo opera la búsqueda heurística por MV, las cadenas MCMC de la IB parten de un árbol al azar (combinación al azar de topología, longitudes de rama y parámetros del modelo) y realiza un cierto número de visitas a árboles concretos. De este modo, los algoritmos MCMC modifican ligeramente una de las variables del árbol inicial (ya sea la posición o longitud de una rama en concreto, o el valor de un parámetro del modelo evolutivo) y evalúan el nuevo árbol. Esto es lo que se conoce como una generación. Tras cada generación, el cambio propuesto se acepta o rechaza en función del valor de la razón entre las probabilidades posteriores de

los estados actual y anterior [36]. Si esta razón es mayor de uno (el nuevo estado es más probable), el cambio se acepta y la cadena MCMC continúa desde este punto. Si por el contrario, la razón es menor de uno, el cambio puede aceptarse o rechazarse en función del valor que toma la razón entre probabilidades posteriores. Recordemos ahora lo que decíamos al hablar de los modelos evolutivos, sobre que los procesos de Markov no tenían memoria. Durante varias generaciones, la cadena MCMC continúa de esta manera y tras haber transcurrido el tiempo suficiente, se espera que la cadena MCMC muestre cada valor de topologías, longitudes de rama y parámetros del modelo un número de veces proporcional a su probabilidad posterior. Es decir, si las relaciones mostradas en la Figura 9.6 tienen una probabilidad posterior de 0.82, la cadena MCMC habrá muestreado esta topología en aproximadamente el 82 % de las generaciones.

Debemos tener en cuenta que la aproximación de la probabilidad posterior será más exacta cuantas más generaciones se evalúen en la cadena de MCMC. Durante las generaciones iniciales de las cadenas MCMC, los árboles suelen tener una probabilidad posterior baja como resultado de haber comenzado a partir de combinaciones aleatorias de topología, longitudes de rama y valores de parámetros. Esto es lo que conocemos como la *fase de burnin* o *calentamiento*. Tras esta fase inicial, las cadenas MCMC alcanzan una fase estacionaria donde los árboles muestreados tienen una probabilidad posterior elevada. Para la estimación del árbol final mediante IB es necesario descartar los árboles del *burnin* y todo el resto de árboles con probabilidad posterior elevada se resumen mediante un árbol consenso de mayoría [36].

Como ocurre con las búsquedas heurísticas de MV, las cadenas MCMC pueden quedar atrapadas en máximos locales. Para evitarlo, la IB utiliza varias cadenas MCMC simultáneamente y además se recomienda hacer cada análisis de IB por duplicado. Los programas de IB como MrBayes [36] generalmente implementan cuatro cadenas MCMC, una cadena “fría” y tres “calientes”. Las cadenas “calientes” elevan la probabilidad posterior de los árboles visitados a una potencia entre 0 y 1, favoreciendo así el salto entre picos locales de probabilidad posterior. Durante el proceso de búsqueda, las cadenas se intercambian, de modo que en cada momento la cadena con mayor probabilidad posterior se comporta como “fría” y el resto como “calientes”. Esto favorece una exploración más exhaustiva del conjunto de árboles posibles, puesto que se facilita el salto entre picos locales de probabilidad posterior elevada y con ello se reduce la posibilidad de que la cadena “fría” quede atrapada en máximos locales. Para el resultado final, sólo se tiene en cuenta los resultados de la cadena “fría”, pues incluye a los árboles con mayor probabilidad posterior entre todos los visitados por las cuatro cadenas.

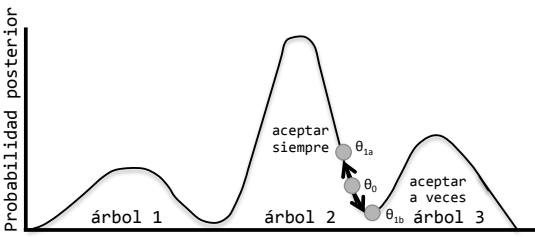
Para saber si las cadenas MCMC de nuestro análisis de IB han explorado de modo eficiente el conjunto de árboles es necesario estudiar la convergencia de las cadenas “frías” entre los dos análisis de IB. La forma más sencilla es observar sus diferencias medias, que en MrBayes conocemos como “*average standard deviation of split frequencies*”. Una diferencia entre las desviaciones estándar medias menor de 0.01 suele indicarnos que los dos análisis de IB han convergido. Sin embargo, es muy recomendable además estudiar la convergencia de las cadenas MCMC a posteriori con herramientas adecuadas desarrolladas a tal efecto, como AWTY (Tabla 9.1) [29].

Los métodos de IB tienen una conexión muy fuerte con la MV, ya que la hipótesis preferida es aquélla con la mayor probabilidad posterior, y éste es un valor que se calcula en función de la verosimilitud. Puesto que la IB se basa directamente en la verosimilitud, comparte sus propiedades de eficiencia y consistencia [18]. Sin embargo, existen diferencias importantes entre MV e IB: la MV trata de encontrar un único árbol (el mejor, el de mayor verosimilitud), mientras que la IB integra un gran número de árboles de probabilidad posterior elevada (con sus topologías y longitudes de rama) [18]. Los árboles con una probabilidad posterior alta generalmente incluyen al mejor árbol de MV, pero también incluyen a un número de árboles cuyo valor de verosimilitud es ligeramente menor. La ventaja de integrar varios árboles es que consideramos la incertidumbre existente en la estimación de la topología, las longitudes

(a) alineamiento

Seq_A	CCTCAGATAC
Seq_B	GGTTAGATAC
Seq_C	GGTCAGATT
Seq_D	GCTCAGACCT

(b) MCMC



(c) Análisis Bayesiano con cuatro cadenas MCMC (x2)

```

1 -- [-2955.775] (-3056.207) (-2997.883) (-2998.675) * [-2911.695] (-3013.663) (-2997.978) (-3016.143)
1000 -- [-2316.864] (-2314.046) (-2322.162) (-2357.051) * [-2377.011] (-2336.706) (-2352.991) [-2301.297]
2000 -- [-2286.754] (-2314.236) (-2298.651) (-2317.366) * [-2325.212] (-2337.705) (-2327.255) [-2285.621]
3000 -- [-2280.581] (-2285.757) (-2303.145) (-2319.858) * [-2305.778] (-2318.057) (-2302.670) [-2285.094]
4000 -- [-2283.396] (-2295.878) (-2312.352) (-2303.106) * [-2286.150] (-2330.199) (-2321.260) [-2298.520]

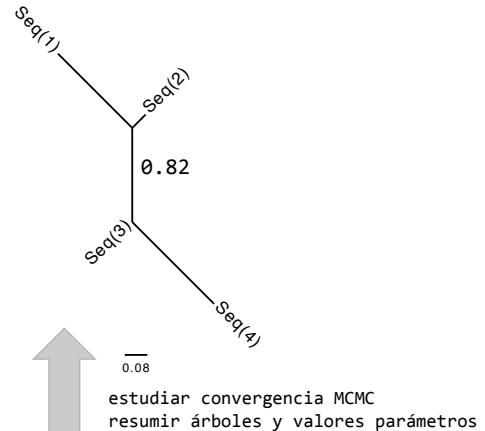
```

Average standard deviation of split frequencies: 0.132583

[...]

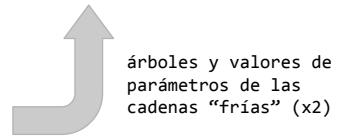
```
10000000 -- [-11441.322] (-11455.888) (-11439.830) (-11462.323) * [-11435.720] (-11458.960) (-11440.275) (-11440.950)
```

Average standard deviation of split frequencies: 0.000584



(c) Resultados de cada generación (x2)

GEN	árbol	par1	par2	parN
1000	(1,2),(3,4)	0,2760	0,1398	0,0005
2000	((1,2),(3,4))	0,2761	0,1398	0,0011
3000	((1,4),(2,4))	0,2779	0,1392	0,0012
4000	((1,2),3,4)	0,2777	0,1399	0,0010
...	...	...	...	...
10000	((1,2),(3,4))	0,2778	0,1390	0,0010



**Figura 9.6:** Inferencia Bayesiana de la filogenia. Las cadenas de *Markov Monte Carlo* (MCMC) progresan hacia árboles más probables gracias a las razones de probabilidad posterior (b). Un análisis de inferencia Bayesiana típico se realiza por duplicado y cada análisis consta de una cadena [fría] y tres (calientes) (c). Tras un número de generaciones determinado, se eliminan las generaciones de la fase burnin y los resultados de las cadenas frías se combinan en el árbol final (d). Con posterioridad, se recomienda revisar la convergencia de las cadenas MCMC.

de rama y los parámetros del modelo evolutivo. Otra diferencia importante entre la IB y el resto de métodos es que resulta necesario especificar la distribución previa de todos los parámetros [18]. Esto nos permite incorporar cierto conocimiento previo en el análisis (por ejemplo, que las transiciones son más probables que las transversiones), pero generalmente no existe tal conocimiento para la mayoría de los parámetros. Por ello, se suelen utilizar distribuciones previas equiprobables (no informativas o *flat priors*) de modo que las diferencias observables de probabilidad posterior dependen principalmente del valor de verosimilitud [36]. A pesar de la fuerte influencia de la verosimilitud en la IB, los árboles de MV e IB pueden ser distintos debido a diferencias durante el proceso de estimación, que es marginal en MV e integrada en IB [18].

## 9.3. Contrastes de hipótesis

### 9.3.1. Contrastes entre modelos evolutivos alternativos

Una de las ventajas de trabajar en un marco evolutivo probabilístico es que podemos recurrir a la estadística para realizar análisis *a posteriori* sobre las filogenias obtenidas. En la Subsección 9.2.2 ya comentamos que existen contrastes, tanto basados en la razón de verosimilitudes (LRT) como en criterios de información (como el AIC), que permiten determinar qué modelo se ajusta mejor a los datos. Además, como ya adelantamos entonces, los contrastes LRT también permiten comparar ciertos modelos que nos ayudan a comprender cómo ha sido el proceso evolutivo que ha dado lugar a las secuencias que observamos. En esta sección introduciremos tres de estos contrastes, que nos permitirán (i) testar la presencia de reloj molecular, (ii) detectar selección positiva y (iii) determinar si dos caracteres han evolucionado de forma correlacionada.

Pero antes explicaremos brevemente cómo realizar el LRT, un tipo de contraste muy recurrente. El LRT requiere que los dos modelos a comparar estén anidados, es decir, que uno de ellos sea un caso especial del otro (el modelo más complejo ha de incluir todos los parámetros del modelo más sencillo). Para realizar el contraste, es necesario calcular la verosimilitud del modelo nulo ( $L_0$ ) y del alternativo ( $L_1$ ), cuya diferencia se mide por medio del estadístico  $D$ , que no es sino el logaritmo de la razón de las verosimilitudes multiplicado por menos dos:

$$D = -2 \ln \frac{L_0}{L_1} = -2 \ln L_0 + 2 \ln L_1 \quad (9.3)$$

Cuando los modelos están anidados, el estadístico  $D$  sigue aproximadamente una distribución chi-cuadrado ( $\chi^2$ ) con grados de libertad igual a la diferencia en el número de parámetros entre ambos modelos. Si el LRT es significativo, se considera que el modelo alternativo (más complejo) se ajusta mejor a los datos; en otras palabras, se rechaza la hipótesis nula.

Para testar si unas secuencias determinadas han evolucionado de acuerdo a una tasa evolutiva constante, es decir si podemos asumir la presencia de un *reloj molecular estricto*, es posible utilizar un contraste de tipo LRT. En primer lugar, calcularemos la verosimilitud del modelo nulo, obtenida al asumir una tasa evolutiva constante (todas las hojas del árbol quedarán alineadas) y la verosimilitud del modelo alternativo donde se permite que las tasas sean variables. En este caso, el test chi-cuadrado tendrá  $s - 2$  grados de libertad, donde  $s$  representa el número de secuencias (el modelo de tasas variables tiene  $s - 1$  parámetros, mientras que el nulo tiene 1 parámetro). Los cálculos de verosimilitud de acuerdo a estos modelos pueden realizarse, por ejemplo, con los programas MEGA, PAML, TREE-PUZZLE e HYPHY (Tabla 9.1). Además de testar la existencia de un reloj molecular global, es posible testar la presencia de relojes moleculares locales, es decir, podemos testar si puede asumirse un reloj molecular en partes concretas de la filogenia.

La importancia de determinar si hay o no constancia de tasas no sólo es útil para comprender cómo ha sido el proceso evolutivo, sino que es importante a la hora de datar árboles filogenéticos. Así, la información contenida en las secuencias puede ser utilizada para estimar tiempos absolutos de divergencia entre linajes (en millones de años). Esto nos permite saber si una separación determinada entre dos linajes ha ocurrido antes o después que otro evento, o si esta separación guarda relación con cierta etapa geológica o climática del planeta. Es por tanto un asunto de gran interés. En el caso de que no podamos asumir la constancia de tasas para todas las secuencias, es posible utilizar lo que conocemos como reloj molecular relajado, que nos permite estimar tiempos de divergencia al mismo tiempo que

acomoda la variación de tasas evolutivas entre linajes [9, 39, 45]. Para conocer más acerca de la datación molecular consultar el texto de Lemey y Posada [27].

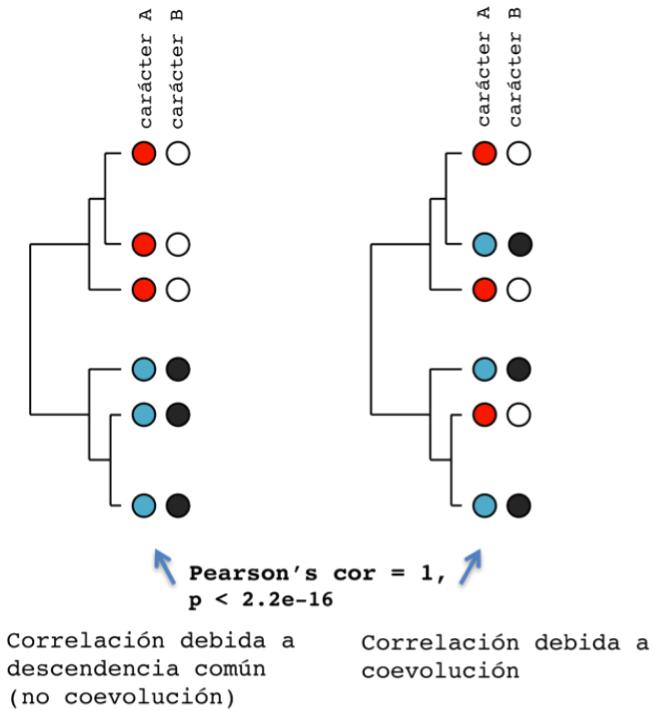
Cuando observamos la morfología de animales o plantas, la *evolución adaptativa* resulta evidente ya que las modificaciones fenotípicas seleccionadas por la evolución tienen un valor directo de adaptación al medio ambiente. A nivel molecular, sin embargo, la evolución adaptativa resulta más difícil de detectar. Aunque hay varias formas de distinguir su huella, la más común se basa en el análisis de secuencias codificantes de proteínas y la naturaleza redundante del código genético. En principio, se trata de comparar el número de cambios sinónimos (sustituciones nucleotídicas que no alteran el aminoácido codificado) por cada sitio sinónimo ( $dS$ ) y el número de cambios no sinónimos por cada sitio no sinónimo ( $dN$ ). Para ello utilizamos modelos de sustitución de codones, que tienen en cuenta factores como las frecuencias de cada codón, la frecuencia relativa de transiciones y transversiones o el sesgo en la composición nucleotídica [47].

La idea subyacente tras la medición de cambios sinónimos y no sinónimos radica en que éstos cambios se acumularán de forma distinta en función de qué fuerza selectiva opere sobre las secuencias. Cuando las secuencias evolucionan de forma neutral, los cambios se acumularán de forma aleatoria, por lo que la razón  $dN/dS$  (también conocida como coeficiente de selección u omega,  $\omega$ ) tenderá a ser 1. En cambio, bajo selección negativa (o purificadora) se espera una proporción mayor de cambios que no alteran el significado y la razón  $dN/dS$  será menor de uno. Si hay selección positiva (o evolución adaptativa), se favorecerán los cambios no sinónimos que aumenten la adaptación y por tanto  $dN/dS$  será mayor de uno. La evolución adaptativa puede observarse con relativa facilidad en genes relacionados con la respuesta inmune, pues requieren de una adaptación constante a los posibles patógenos que deben combatir y con los que coevolucionan (lo que se conoce como una *carrera armamentista* o *hipótesis de la Reina de Corazones* o *Red Queen*).

La huella de la selección positiva es con frecuencia sutil y desaparece rápidamente. En primer lugar, la evolución adaptativa suele actuar de forma episódica (en un momento y circunstancias determinadas) y después la selección se torna purificadora con el fin de mantener dicha adaptación. Por tanto, tras un episodio de evolución adaptativa, la razón  $dN/dS$  puede volver rápidamente a ser menor de uno. En segundo lugar, la selección positiva no actúa sobre toda la proteína, sino que generalmente afecta sólo a posiciones concretas que tienen importancia funcional. Por tanto, un valor de  $\omega$  puede ser menor de uno si lo promediamos para todas las posiciones, pero al mismo tiempo puede enmascarar ciertas posiciones donde omega es mayor de uno. Para tratar de resolver estas dificultades existen modelos sofisticados que nos permiten enfocar el cálculo de  $\omega$  en regiones concretas del árbol (se conocen como *modelos de ramas* o *branch models*), en las distintas posiciones de las proteínas (*modelos de sitios* o *site models*), o en ambos simultáneamente (*modelos de ramas y sitios* o *branch-site models*). Los programas más populares para realizar este tipo de análisis (y otros relacionados) son PAML y HYPHY (Tabla 9.1). El manual de usuario de PAML explica en detalle cómo realizar distintos contrastes e incluye ejemplos que nos permitirán reproducir los resultados de trabajos originales. Para conocer más acerca de los métodos que permiten detectar la selección adaptativa leer el texto de Kosakovsky, Pond et al. [25].

En tercer lugar, además de evaluar las hipótesis de reloj molecular y de evolución adaptativa, es posible utilizar contrastes de tipo LRT para determinar si dos caracteres (sean morfológicos o moleculares) han evolucionado de forma correlacionada. La *evolución correlacionada* también se conoce como coevolución o evolución dependiente. Con frecuencia, los análisis de correlación clásicos (como el coeficiente de correlación de Pearson) no pueden aplicarse directamente en sistemas biológicos. Estos métodos asumen la independencia de los datos, pero las especies y sus atributos no son independientes, sino que están relacionados mediante su filogenia [13]. Muchas de las características que comparten los seres vivos se deben a descendencia común de éstos, por lo que la presencia conjunta de ciertas características no implica necesariamente que éstas hayan coevolucionado, sino que puede sencillamente deberse a un

origen común. En el ejemplo de la Figura 9.7 se muestran dos escenarios evolutivos completamente distintos en los que en ambos el coeficiente de Pearson resulta en una correlación alta. Sin embargo, el panel izquierdo refleja que esta correlación es fruto de descendencia común, mientras que el panel derecho refleja claramente la coevolución de dos caracteres con independencia de su filogenia. Este tipo de análisis que tiene en cuenta la historia evolutiva, puede realizarse con BayesTraits (Tabla 9.1) y se basa, nuevamente, en contraste de modelos evolutivos mediante un LRT.



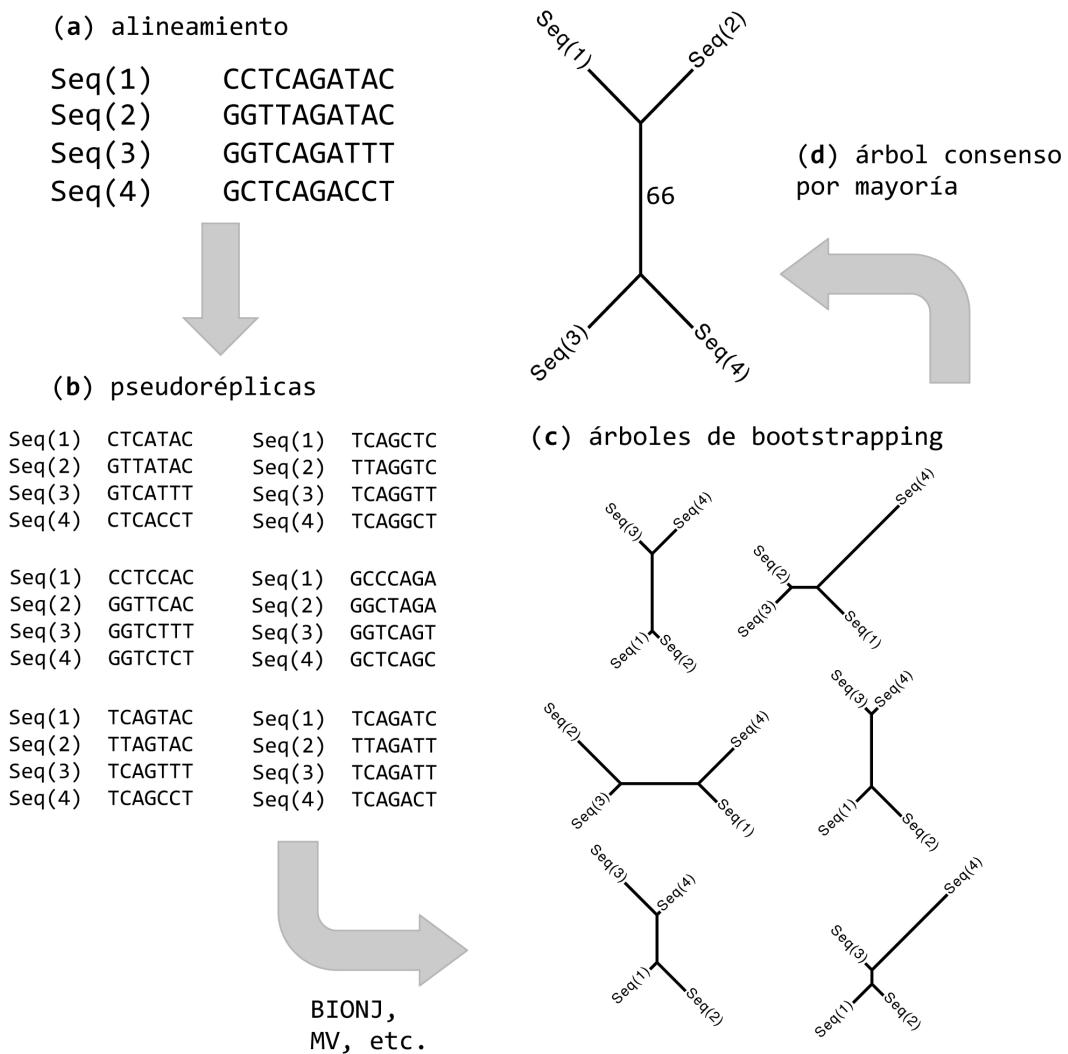
**Figura 9.7:** Cálculo de correlaciones en un contexto evolutivo. Se muestra cómo dos escenarios evolutivos completamente diferentes pueden dar lugar a un mismo valor de correlación de Pearson, si bien en un caso la correlación refleja descendencia común (a) y en otro coevolución (b).

### 9.3.2. Robustez de un árbol filogenético y contraste de árboles alternativos

Uno de los aspectos más importantes del análisis filogenético es el de poder determinar cuán fiable es el árbol que hemos obtenido, y si existen hipótesis alternativas con una fiabilidad similar. La inferencia filogenética mediante métodos de MV produce un único árbol, el árbol que con mayor verosimilitud ha dado lugar a los datos observados. Por esta razón, no conocemos de manera directa si existe otro árbol con una verosimilitud similar no significativamente diferente desde el punto de vista estadístico.

De cualquier modo, la reconstrucción filogenética por MV no define intervalos de confianza para los árboles estimados y es necesario utilizar métodos alternativos. La aproximación más corriente es utilizar el bootstrapping no paramétrico [12], que nos informa acerca de la estabilidad de las relaciones filogenéticas del árbol de MV obtenido (aunque no puede informar de la presencia de otros árboles subóptimos). El *bootstrapping no paramétrico* es una técnica estadística de remuestreo con reemplazamiento, que funciona de la siguiente manera: a partir de nuestro alineamiento original (Figura 9.8(a)) se generan pseudoréplicas muestreando aleatoriamente una proporción concreta de las columnas (sitios) del alineamiento original (Figura 9.8(b)). En este muestreo con reemplazamiento se escogen columnas

al azar para generar las pseudorélicas, pudiendo incluir una misma columna varias veces mientras que otras nunca se incluirán. Una vez que se han generado un cierto número de pseudorélicas (por ejemplo, 1000), se infiere un árbol filogenético con un método particular (por ejemplo, MV) para cada una de ellas, de manera que obtenemos 1000 árboles de MV (Figura 9.8(c)). El conjunto de árboles obtenidos a partir de las pseudorélicas se resume en un árbol final que conocemos como árbol de consenso por mayoría (Figura 9.8(d)). Este árbol incluirá un número (típicamente expresado en %) para cada uno de los nodos del árbol en función del número de veces que se encuentre esta relación en particular en el conjunto de los 1000 árboles de *bootstrap*. Generalmente se considera que un valor igual o mayor al 70 % es un soporte elevado para las relaciones de parentesco implicadas ([17]).



**Figura 9.8:** Bootstrapping no paramétrico para estimar la robustez de un árbol. El alineamiento original (a) se remuestra al azar y con reemplazamiento N veces, obteniendo N alineamientos con la misma longitud (b). Para cada uno de estos alineamientos se estima el mejor árbol, ya sea por MV u otros métodos, obteniéndose N árboles (c). Finalmente, se analizan los N árboles y se genera un árbol consenso (d).

A diferencia de los métodos filogenéticos de MV, el resultado final de la IB tiene en cuenta un número elevado árboles, lo que nos permite realizar una estima de la robustez de nuestro árbol final gracias al

proceso de búsqueda por MCMC. Para ello, hacemos uso de los árboles obtenidos en las generaciones tras la convergencia (la fase estacionaria), que se analizan *a posteriori* para calcular la frecuencia con la que las relaciones de parentesco ha sido obtenidas en el conjunto de árboles. Esta frecuencia se expresa mediante valores de probabilidad posterior para cada uno de los nodos, de modo análogo a la frecuencia calculada a partir de las réplicas de *bootstrapping* no paramétrico. En el caso de probabilidades posteriores, se considera que un nodo en particular es fiable cuando su probabilidad posterior es superior a 0.95.

Además de estimar la fiabilidad de los árboles filogenéticos, es posible utilizar test estadísticos para saber si nuestros datos, además de favorecer una hipótesis filogenética concreta (nuestro árbol reconstruido) podrían soportar o rechazar otras hipótesis filogenéticas, como las propuestas en estudios previos. Este tipo de análisis son muy interesantes, y pueden aplicarse, entre otras cosas, para conocer si nuestros datos de secuencia apoyan o rechazan determinadas relaciones de parentesco derivadas del análisis de datos morfológicos o de otro tipo.

Existen varios métodos para comparar la topología de árboles filogenéticos. Los más empleados utilizan el remuestreo de caracteres y la comparación de las verosimilitudes de dos árboles alternativos. Sin embargo, dado que las topologías no son hipótesis anidadas (un árbol generalmente no representa un caso particular de otro árbol), las diferencias de verosimilitud entre árboles no pueden aproximarse mediante una distribución chi-cuadrado y no es posible utilizar un LRT como hemos explicado hasta ahora.

Una posibilidad para comparar hipótesis no anidadas mediante LRT es utilizar el *bootstrapping* paramétrico, que nos permite determinar la distribución nula del estadístico D [19]. El *bootstrapping* no paramétrico, también conocido como simulación de Monte Carlo, consiste en generar réplicas de los datos y estimar mediante un método como la MV un conjunto de árboles (con su topología, longitudes de rama y parámetros del modelo) que formarán la distribución nula. En este caso, el LRT será significativo si las diferencia entre las verosimilitudes de los árboles de la distribución nula y la hipótesis alternativa superan el 5 % de los casos [19]. El *bootstrapping* paramétrico puede resultar computacionalmente muy intensivo, especialmente para grandes conjuntos de datos. Por esta razón, la comparación de topologías alternativas suele realizarse mediante contrastes de topologías basados en *bootstrapping* no paramétrico.

Los test de topologías alternativas más populares son tres. El *contraste de Kishino-Hasegawa* (KH) [24] permite comparar la diferencia de verosimilitudes entre dos árboles seleccionados *a priori*. Este contraste estadístico sólo es válido si los árboles se escogen previamente y no se derivan del mismo alineamiento que se utilizará para compararlos posteriormente. El *contraste de Shimodaira-Hasegawa* (SH) [42] permite comparar un conjunto de árboles seleccionados *a posteriori*, lo cual nos permite incluir el árbol de MV obtenido a partir de nuestro alineamiento. Sin embargo, el contraste SH es muy conservador y sus resultados son muy dependientes del número de árboles considerados. Existe un tercer tipo de contraste conocido como *test aproximadamente no sesgado* (AU, del inglés *approximately unbiased test*) que está menos sesgado que los contrastes desarrollados con anterioridad (KH, SH), y por ello es el más utilizado actualmente [41]. El contraste AU utiliza una aproximación de *bootstrap* multiescala para controlar el error de tipo I (rechazar inapropiadamente la hipótesis nula). El procedimiento de *bootstrap* multiescala consiste en generar pseudoréplicas de *bootstrap* como en el caso de *bootstrapping* no paramétrico, pero en este caso las pseudoréplicas son de longitud variable, de modo que ciertas pseudoréplicas serán de longitud mayor y otras de longitud menor que el alineamiento original. Tras contar el número de veces que una hipótesis filogenética es apoyada por las pseudoréplicas, se obtienen valores de *bootstrap* para cada conjunto de pseudoréplicas de una determinada longitud. El valor de probabilidad del test AU se calcula a partir de las diferencias en estos valores de *bootstrap* en función de las longitudes de las pseudoréplicas. Para conocer más acerca de los test de topologías consultar el

texto de Heiko Schmidt [40].

## 9.4. Reconstrucción de estados ancestrales

Además de permitirnos estudiar las relaciones de parentesco entre organismos (o genes) y comprender el proceso evolutivo que ha dado lugar a éstos, la inferencia filogenética también nos permite reconstruir el pasado. Si bien los fósiles pueden darnos información muy valiosa acerca de las características externas y el modo de vida de especies extintas, la información molecular de éstas es generalmente inexistente (aunque hemos sido capaces de secuenciar el genoma del Neandertal).

Existen diversos métodos que nos permiten inferir cuáles eran los *estados ancestrales* de un carácter, ya sea morfológico, molecular (por ejemplo, los aminoácidos) o incluso una variable cuantitativa [31]. Los métodos basados en parsimonia prefieren las reconstrucciones más sencillas (que requieran un número menor de cambios). Estos métodos tienen ciertas limitaciones, por lo que generalmente preferimos utilizar métodos probabilísticos que reconstruyen los estados ancestrales de una forma estadística. La MV permite hallar cuál era el estado ancestral más probable, mientras que los métodos Bayesianos calculan la probabilidad posterior para cada posible estado de dicho carácter. Los programas más robustos y prácticos para reconstruir estados ancestrales mediante métodos probabilísticos (MV e IB) son MrBayes, BayesTraits y PAML (Tabla 9.1). Existen guías y tutoriales que explican detalladamente cómo utilizar estos programas.

A continuación, presentaremos un ejemplo que sirve para ilustrar cómo resolver enigmas del pasado mediante la reconstrucción de los estados ancestrales de una proteína. Las opsinas son una proteínas que contienen como grupo prostético el 11-*cis*-retinal que se excita cuando capta un fotón. El cambio conformacional consecuente en la proteína produce una señal que se traduce a través de las proteínas G y finalmente llega a nuestro cerebro, confiriéndonos la capacidad de ver. Existen distintas opsinas, y cada una es sensible a una longitud de onda (color) característica. Es sabido que los mamíferos poseemos una excelente visión nocturna (los humanos no tanto, en realidad), y se ha propuesto que esta cualidad, junto a las de ser homeotermos y tener un buen abrigo (el pelo), habría permitido a nuestros ancestros llevar un estilo de vida nocturno durante el reinado de los dinosaurios en el Jurásico, que se creían animales principalmente diurnos. No obstante, un estudio reconstruyó la secuencia ancestral de las opsinas de los arcosaurios (grupo que engloba, entre otros, cocodrilos, dinosaurios y aves) y demostró, tras ensayar su actividad molecular *in vitro*, que la proteína reconstruida mediante métodos filogenéticos era completamente funcional (a pesar de su enorme antigüedad de 420 millones de años) y que además mostraba una elevada fotosensibilidad similar a la de las opsinas de los mamíferos actuales [6]. Este estudio puso en duda la hipótesis de que los dinosaurios eran diurnos, puesto que al menos potencialmente serían capaces de ver en la oscuridad. Éste y otros ejemplos de la utilidad de la reconstrucción de estados ancestrales pueden encontrarse en el artículo de Thornton [46].

## 9.5. Guía rápida de reconstrucción filogenética

En este capítulo hemos presentado los fundamentos de la reconstrucción filogenética con cierto detalle. A continuación ofrecemos una guía rápida (pero minuciosa) para reconstruir con fiabilidad un árbol filogenético.

1. *Selección de los datos.* La elección de qué secuencias y de qué especies hemos de incluir en el análisis depende del tipo de pregunta que queramos responder. Además, este paso suele ser iterativo, pues la obtención de un árbol nos puede sugerir cómo mejorar el muestreo de caracteres

o secuencias, por ejemplo, buscando un grupo externo (o *outgroup*) más adecuado o incluyendo o excluyendo ciertas secuencias.

2. El primer paso es la construcción de un *alineamiento múltiple de secuencias*, ya sean nucleótidos o aminoácidos. En el caso del alineamiento de nucleótidos correspondientes a secuencias que codifican proteínas, es recomendable utilizar herramientas como TranslatorX (Tabla 9.1) que determinan el alineamiento de nucleótidos en función del alineamiento de aminoácidos correspondiente. Esto lo hacemos así porque es más fácil alinear secuencias de aminoácidos que secuencias de nucleótidos, y además se respeta la integridad de los codones.
3. Seguidamente debemos asegurarnos de la *calidad de nuestro alineamiento*. Errores de secuenciación o de anotación incorrecta de los genes puede producir secuencias con inserciones o delecciones considerables. Además, es necesario eliminar las posiciones de homología posicional dudosa (regiones muy variables o con muchos gaps), para lo cual es conveniente utilizar herramientas como GBlocks o TrimAl (Tabla 9.1). De este modo reducimos el problema de la falsa homología, y hasta cierto punto el de saturación y homoplásia, mejorando el resultado de la inferencia filogenética.
4. Antes de proceder a inferir el árbol filogenético, normalmente debemos *determinar el modelo de evolución que se ajusta mejor a nuestros datos*. Los programas jModelTest y ProtTest implementan la selección de modelos de nucleótidos y aminoácidos (respectivamente), y ModelGenerator es capaz de trabajar con ambos tipos (Tabla 9.1). Estos programas ordenan los distintos modelos en función de nuestro criterio de selección (generalmente, AIC). A partir de esta lista, elegiremos aquel modelo que mejor se ajuste a nuestros datos (o sus particiones) y que al mismo tiempo sea implementado en el programa de reconstrucción que vayamos a utilizar (no todos los modelos están disponibles en los programas de reconstrucción filogenética).
5. *Los métodos de distancias* como BIONJ nos permiten obtener un *árbol de forma rápida*, que nos puede servir para hacernos una idea de las relaciones filogenéticas entre nuestras secuencias. Así, estos árboles preliminares nos permiten identificar rápidamente la existencia de ramas largas, una de las fuentes principales de artefactos filogenéticos (ver también punto 6). Para este tipo de reconstrucciones podemos servirnos de los programas SeaView o MEGA (Tabla 9.1), que integran en un entorno gráfico la gestión de alineamientos múltiples y la reconstrucción de árboles por BIONJ y otros métodos (aunque no todos).
6. *Para obtener un árbol más fiable, utilizaremos métodos probabilísticos* como MV o IB. Para la reconstrucción por MV recomendamos PhyML y RAxML (Tabla 9.1). El primero puede ejecutarse en el servidor web de PhyML, desde la consola o puede ser utilizado desde el programa SeaView. La IB generalmente demanda más esfuerzo de computación y es más lenta, pero se puede paralelizar. El programa más popular de IB es MrBayes, si bien existen otras alternativas como PhyloBayes, que implementan modelos evolutivos más complejos (Tabla 9.1). MrBayes utiliza un tipo de fichero llamado nexus, donde además del alineamiento, se incluye un bloque con las instrucciones para el análisis (modelo evolutivo, número de análisis paralelos, número de cadenas MCMC, número de generaciones, etc.). El servidor web MrBayes Web Form puede servir de ayuda para preparar estos ficheros (Tabla 9.1).
7. Para *visualizar los árboles* obtenidos podemos utilizar los visores integrados de los programas SeaView o MEGA, aunque son limitados en sus funciones. De forma alternativa, podemos recurrir al programa FigTree, que, además de ser más versátil, permite preparar figuras atractivas para su publicación (Tabla 9.1).
8. Frecuentemente, la inspección de los árboles nos sugiere que debemos refinar el conjunto de datos seleccionados (vuelta al punto “0”). Por ejemplo, cuando hay secuencias que han divergido mucho

más que las demás, sus ramas extremadamente largas pueden producir el fenómeno conocido como atracción de ramas largas, un tipo de artefacto filogenético bien caracterizado. Si fuera posible es recomendable evitar la inclusión de estas secuencias.

9. Normalmente el siguiente paso es el de *determinar cuán creíble es el árbol* que hemos obtenido. Si utilizamos MV, lo más habitual es realizar un *bootstrapping* no paramétrico, que está implementado en los programas de inferencia filogenética por MV. Podemos elegir 1000 réplicas, y los nodos que obtengan un soporte mayor de 70 % se considerarán fiables. Si la reconstrucción la hemos hecho con un método de IB, el soporte de los nodos se mide mediante probabilidades posteriores, que ya vienen implícitas en los resultados de la IB. Normalmente se consideran fiables aquellos nodos cuya probabilidad posterior es mayor de 0.95.
10. Finalmente, cuando queremos publicar un árbol filogenético, lo más recomendable es *combinar dos métodos probabilísticos*, MV e IB, y representar conjuntamente los resultados de ambos, indicando los valores de *bootstrap* y probabilidad posterior junto a cada nodo del árbol (o sobre los nodos de mayor interés).

## 9.6. Programas recomendados

Programa	Propósito	Modo de uso
ALTER	Conversión de formatos de alineamientos.	web
AWTY	Estudio de la convergencia de las cadenas MCMC en análisis de IB.	web
BayesTraits	MV e inferencia Bayesiana de estados ancestrales, coevolución y otros modelos evolutivos, tanto para caracteres discretos como continuos.	local
CONSEL	Contraste de hipótesis de topologías alternativas (test KH, SH, AU).	local
FigTree	Visualización y edición de árboles filogenéticos.	local
GARLI	Reconstrucción filogenética por MV.	
GBlocks	Limpieza de alineamientos.	web, local
HYPHY	Contraste de hipótesis evolutivas. Muy versátil.	local, web
JalView	Visor y editor de alineamientos múltiples.	web, local
jModelTest	Selección de modelos de evolución de nucleótidos.	local
MEGA	Múltiples posibilidades: reconstrucción filogenética, contrastes de hipótesis.	local
Mesquite	Evolución de caracteres, edición de árboles.	local
ModelGenerator	Selección de modelos de evolución.	local
MrBayes	Reconstrucción filogenética con métodos Bayesianos, reconstrucción de estados ancestrales, modelos evolutivos de aminoácidos, nucleótidos y codones. Posibilidad de combinar datos moleculares y morfológicos.	local
MrBayes Web Form	Generación de archivos en formato nexus que incluyen comandos específicos de MrBayes.	web
PAML	Contraste de hipótesis evolutivas: selección natural, reloj molecular, evolución adaptativa, simulaciones.	local
PartitionFinder	Selección de esquemas de partición.	local
PAUP	Múltiples y diversos métodos y test.	local
Phylip	Paquete de programas con múltiples funcionalidades.	web, local
PhyML	Reconstrucción filogenética por MV. Modelos de evolución de nucleótidos, aminoácidos y codones (codonPhyML).	web, local, SeaView
PhyloBayes	Reconstrucción filogenética con métodos Bayesianos. Implementa modelos evolutivos complejos que pueden acomodar fluctuaciones en los parámetros del modelo a lo largo del alineamiento.	local
ProtTest	Selección de modelos de evolución de proteínas.	web, local
ReadSeq	Conversión de formatos de alineamientos.	web, local
SeaView	Alineamiento, Neighbour-joining, Máxima parsimonia, Phyml, bootstrapping.	local

TranslatorX	Alineamiento de secuencias nucleotídicas guiado por el correspondiente alineamiento de proteínas.	web, local
Tree-Puzzle	Reconstrucción filogenética por MV utilizando el método quartet-puzzling para la búsqueda de árboles. Incluye modelo de reloj molecular.	local
TrimAl	Limpieza de alineamientos.	web, local

**Tabla 9.1:** Algunos programas útiles para la inferencia de árboles filogenéticos y el estudio evolución molecular. Ésta no pretende ser una lista exhaustiva, sino una recomendación personal de los autores de este capítulo.

## 9.7. Bibliografía

- [1] F. Abascal, R. Zardoya, and D. Posada. Prottest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–2105, 2005.
- [2] L. Bromham. Why do species vary in their rate of molecular evolution? *Biology Letters*, 5(3):401–404, 2009.
- [3] D. Bryant, N. Galtier, and M.-A. Poursat. *Likelihood calculation in molecular phylogenetics*, pages 33–58. Oxford University Press, Oxford, New York, 2005.
- [4] S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–3, 2009.
- [5] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552, 2000.
- [6] B. S. W. Chang and M. J. Donoghue. Recreating ancestral proteins. *Trends in Ecology & Evolution*, 15(3):109–114, 2000.
- [7] M. D. Crisp and L. G. Cook. Do early branching lineages signify ancestral traits? *Trends in Ecology & Evolution*, 20(3):122–128, 2005.
- [8] C. Darwin. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. John Murray, London, 1859.
- [9] A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88, 2006.
- [10] A. W. F. Edwards and L. L. Cavalli-Sforza. *Reconstruction of evolutionary trees*, pages 67–76. Systematics Association, London, 1964.
- [11] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.
- [12] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.
- [13] J. Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.
- [14] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, 2004.
- [15] O. Gascuel. Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695, 1997.
- [16] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [17] D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2):182–192, 1993.
- [18] M. Holder and P. O. Lewis. Phylogeny estimation: traditional and bayesian approaches. *Nature Reviews Genetics*, 4:275–284, 2003.
- [19] J. P. Huelsenbeck and K. A. Crandall. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28(1):437–466, 1997.
- [20] L. S. Jermiin, V. Jayaswal, F. Ababneh, and J. Robinson. *Phylogenetic model evaluation*, volume 452 of *Methods in Molecular Biology*, pages 331–364. Springer Verlag, Notowa, 2008.
- [21] T. H. Jukes and C. R. Cantor. *Evolution of protein molecules*, volume III, pages 21–132. Academic Press, New York, 1969.
- [22] T. Keane, C. Creevey, M. Pentony, T. Naughton, and J. McInerney. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology*, 6(1):29, 2006.
- [23] M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

- [24] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29(2):170–179, 1989.
- [25] S. L. Kosakovsky Pond, A. F. Y. Poon, and W. Frost, Simon D. *Estimating selection pressures on alignments of coding sequences*, pages 419–490. Cambridge University Press, New York, second edition edition, 2009.
- [26] R. Lanfear, B. Calcott, S. Y. Ho, and S. Guindon. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol*, 29(6):1695–701, 2012.
- [27] P. Lemey and D. Posada. *Molecular clock analysis*, pages 362–377. Cambridge University Press, New York, second edition edition, 2009.
- [28] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.
- [29] J. A. Nylander, J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford. Awty (are we there yet?): A system for graphical exploration of mcmc convergence in bayesian phylogenetics. *Bioinformatics*, 24(4):581–583, 2008.
- [30] K. E. Omland, L. G. Cook, and M. D. Crisp. Tree thinking for all biology: the problem with reading phylogenies as ladders of progress. *BioEssays*, 30(9):854–867, 2008.
- [31] M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999.
- [32] D. Posada. jmodeltest: phylogenetic model averaging. *Molecular Biology and Evolution*, 25(7):1253–1256, 2008.
- [33] D. Posada. *Selecting models of evolution*, pages 345–361. Cambridge University Press, New York, second edition edition, 2009.
- [34] D. Posada and T. R. Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol*, 53(5):793–808, 2004.
- [35] J. H. Reeves. Heterogeneity in the subsitution process of amino acid sites of proteins coded by mitochondrial dna. *Journal of Molecular Evolution*, 35:17–31, 1992.
- [36] F. Ronquist, P. van der Mark, and J. P. Huelsenbeck. *Bayesian phylogenetic analysis using MrBayes*, pages 210–266. Cambridge University Press, New York, second edition edition, 2009.
- [37] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5):1073–1095, 1993.
- [38] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.
- [39] M. J. Sanderson. Nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, 14:1218–1231, 1997.
- [40] H. Schmidt. *Testing tree topologies*, pages 381–403. Cambridge University Press, New York, second edition edition, 2009.
- [41] H. Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51:592–508, 2002.
- [42] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16:1114–1116, 1999.
- [43] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. *Phylogenetic inference*, pages 407–514. Sinauer Associates, Sunderland, Massachusetts, 1996.
- [44] S. Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.
- [45] J. L. Thorne, H. Kishino, and I. S. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647–1657, 1998.
- [46] J. W. Thornton. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Review Genetics*, 5:366–375, 2004.
- [47] Z. Yang. *Computational molecular evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, New York, 2006.

- [48] E. Zuckerkandl and L. Pauling. *Molecular disease, evolution, and genetic heterogeneity*, pages 189–225. Academic Press, New York, 1962.