

# Filogenia molecular

---

## Resumen

La filogenia molecular es la rama de la filogenia que analiza las diferencias moleculares hereditarias en las secuencias de ADN, ARN y proteínas para obtener información sobre las relaciones evolutivas de un organismo. El resultado de un análisis filogenético molecular se expresa en un árbol filogenético.

# Índice general

<b>I</b>	<b>Introducción a la filogenia: principios y conceptos</b>	<b>2</b>
I.1	Conceptos básicos	3
I.2	Politomías	4
I.3	Tipos de árboles filogenéticos	5
I.4	Inferencia filogenética	6
I.5	Homología	6
I.5.1	Tipos de homología	6
I.5.2	Homoplasia	7
I.5.3	Agrupamientos	8
I.5.4	Fenotipo vs moléculas	8
<b>II</b>	<b>Alineamiento de secuencias</b>	<b>10</b>
II.1	Tipos de caracteres	10
II.2	Ponderación de los caracteres	10
II.2.1	Ponderación <i>a priori</i>	11
II.2.2	Ponderación <i>a posteriori</i>	11
II.2.3	Secuencias de ADN	11
II.3	Polaridad de los caracteres	12
II.4	Homología de los caracteres moleculares	12
II.4.1	Aplicación del concepto de homología a los genes: alineamiento de secuencias	12
II.4.2	Decidir el mejor alineamiento	13
<b>III</b>	<b>Modelos de evolución</b>	<b>14</b>
III.1	Modelos frecuentes	15
III.2	Heterogeneidad de tasas de sustitución	15
III.3	Elegir el mejor modelo	16
III.3.1	Selección de esquemas de partición	16
<b>IV</b>	<b>Métodos filogenéticos de inferencia</b>	<b>18</b>
IV.1	Búsqueda de árboles	18
IV.2	Árboles de consenso	19
IV.3	Medidas de soporte: confianza en el árbol	22
<b>V</b>	<b>Máxima parsimonia - Cladística</b>	<b>24</b>
V.1	Construir el árbol	24
V.2	Enraizar el árbol	26
V.3	Valoración de árboles	27

# Capítulo I

## Introducción a la filogenia: principios y conceptos

La filogenia es la determinación de la historia evolutiva de los organismos. Así, la filogenética es el estudio de: (i) la filogenia mediante el uso de árboles filogenéticos de los distintos organismos y (ii) las relaciones entre ellos. Ha habido varias iniciativas a lo largo de la historia (como [ToL Web](#)) que han intentado lograr crear árboles de todas las especies, cuyo número se estima por lo bajo que ronda los 3-5 millones. Estas estimaciones sobre la biodiversidad se realizan a partir de el número de grupos clave como los escarabajos, que son el grupo con mayor número de spp., en un lugar.

La filogenia es una disciplina muy consolidada; desde sus inicios, hace aproximadamente 200 años, sus representaciones no han cambiado mucho. La filogenia trabaja con árboles evolutivos, que son las **representaciones gráficas (patrones) de las relaciones ancestro-descendientes (relaciones históricas de parentescos) entre elementos**, que pueden ser especies, secuencias de genes, etc. Entender este patrón es esencial para realizar estudios comparativos de cualquier tipo, porque existen **dependencias estadísticas entre los elementos que comparten ancestros comunes**. Conforme pasa el tiempo, se van aplicando diferentes y nuevos modelos evolutivos y se van depurando. En última instancia, se se obtiene una mejor aproximación cuantos más datos se añadan (tanto más especies como más secuencias) a los modelos.

La filogenia sirve, entre otros, para:

- Evolución de los seres vivos
- Genómica
- Ingeniería genética
- Farmacia
- Epidemiología
- Biología de la conservación
- Control de plagas

## ■ Lingüística

La filogenética puede ser estudiada de diversas maneras. A menudo ha sido estudiada utilizando registros fósiles, que contienen información sobre la morfología de los antepasados de las especies actuales y la cronología de sus divergencias. Esto permite datar las filogenias. Sin embargo, el uso de registros fósiles presenta muchas limitaciones: pueden estar disponibles sólo para determinadas especies, los datos existentes de fósiles pueden estar fragmentados, la recolección de datos está limitada por la abundancia, hábitat, rango geográfico y otros factores; y las descripciones de los rasgos morfológicos son a menudo ambiguas (múltiples factores genéticos). Por todo esto, utilizar registros fósiles para determinar relaciones filogenéticas puede producir sesgos. Además, los fósiles de microorganismos son prácticamente inexistentes imposibilitando el uso de este enfoque. Afortunadamente, los datos moleculares que están en la forma de secuencias de ADN o de proteínas pueden ser también muy útiles para proporcionar una perspectiva de la evolución de los organismos, como el ARN 16S. Debido a que los genes son el medio para registrar las mutaciones acumuladas, éstos pueden servir como "fósiles moleculares". A través del análisis comparativo de secuencias de ADN de una serie de organismos relacionados, la historia evolutiva de los genes e incluso de los organismos puede ser revelada. La ventaja de utilización de datos moleculares es que son más numerosos que los registros fósiles y más fáciles de obtener. Además, no hay ningún sesgo de muestreo, como el que hay en los registros fósiles reales. Por tanto, es posible construir árboles filogenéticos más precisos y robustos utilizando datos moleculares.

Las filogenias, al ser una reconstrucción de la evolución de los caracteres, se construyen a partir de un registro/evidencias indirecto/as del proceso evolutivo. Por tanto, se deben realizar test de homología comprobar que los caracteres son comparables entre sí al compartir un origen común (homologías) y discernirlas de las homoplasias. De esa forma se obtiene información para construir clasificaciones y hacer predicciones dentro de un marco temporal cuando es posible obtenerlo.

## I.1. Conceptos básicos

Los árboles filogenéticos suelen ser binarios, estando compuestos por **nodos externos o terminales** y **nodos internos** unidos por **ramas** que parten de una **raíz**. A través de las diferentes ramas se van reconstruyendo las relaciones entre las especies. Los **nodos internos son hipótesis evolutivas de posibles ancestros comunes** de los cuales normalmente faltan datos para confirmar o descartar la teoría. En las distintas ramas se pueden representar la transformación de caracteres que aparecen a nivel genético y que se transmiten por herencia.

Se denominan **grupos hermanos** a los nodos terminales que parten de un mismo nodo interno, es decir, dos taxones que compartan un ancestro común no compartido por ningún otro taxón. El **grupo externo (outgroup)** es aquel que se encuentra más alejado y parte de una rama distinta desde la raíz. Normalmente, este outgroup se elige de forma consciente para poder colocar la raíz donde se estima correcto. Todas las especies que se desarrollan desde una rama de la raíz se denomina **grupo interno o ingroup**.

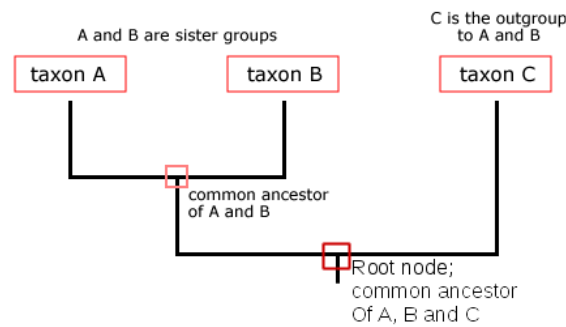


Figura I.1: Partes de un árbol filogenético.

Los árboles filogenéticos se pueden representar sin enraizar o enraizado. Un árbol filogenético sin raíz no asume conocimiento de un ancestro común, solo posiciones de los taxones para mostrar sus relaciones relativas (no hay dirección de un camino evolutivo). Para definir la dirección de la evolución se necesita un árbol filogenético con raíz donde todas las secuencias bajo estudio tienen un ancestro o nodo raíz común (más informativo). Mientras que los árboles filogenéticos se centran en las relaciones evolutivas entre diferentes especies, las redes haplotípicas son representaciones gráficas sobre las relaciones evolutivas entre las diferentes poblaciones.

A la hora de visualización, hay varias formas de representar los árboles filogenéticos. Los distintos elementos no tienen un orden concreto; da igual si en un árbol los nodos terminales están en distinto orden mientras que las ramas sigan el mismo camino. En general, se suelen poner los nodos terminales de manera que sea más fácil de leer a simple vista.

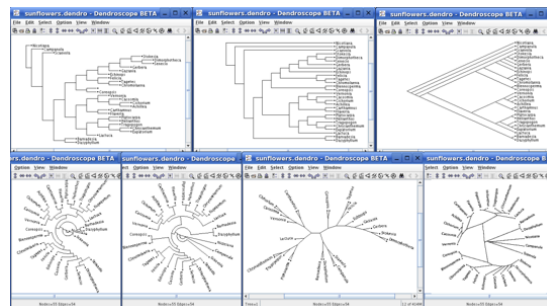


Figura I.2: Distintas representaciones de los árboles filogenéticos.

## I.2. Politomías

La topología es la forma en que se ramifica un árbol. Cuando todas las ramas se bifurcan en un árbol filogenético, éstas son denominadas como una **dicotomía**. Por el contrario, si una rama tiene más de dos descendientes, entonces se denomina **politomía**.

Los árboles filogenéticos se consideran resueltos cuando de un nodo interno salen las distintas terminales. En la mayoría de casos, los árboles son no resueltos y tienen politomías, es decir, que desde un nodo interno no se sabe cómo han avanzado las especies. A partir de ahí solo se pueden añadir más datos, pintar uniones con un

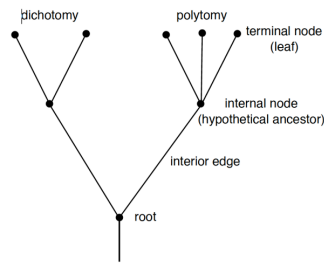


Figura I.3: Diferencia entre dicotomía y politomía.

bootstrap bajo (es decir, un bajo soporte de esa bifurcación), o justificarlo como que están en el momento de especiación. Dentro de las hipótesis filogenéticas siempre hay más de una solución (se producen varios árboles igualmente óptimos), así que el árbol final se debe elegir. Un árbol de consenso puede ser construido mostrando las porciones de bifurcación resueltas comúnmente y colapsando aquellas que no concuerdan entre los árboles. En un árbol de consenso estricto, todos los nodos en conflicto son colapsados.

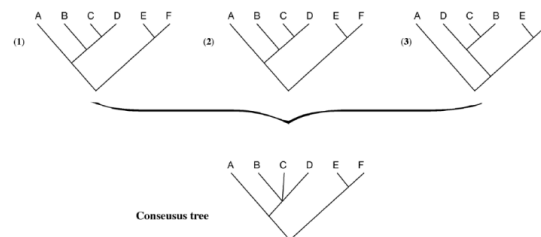


Figura I.4: Árbol de consenso.

### I.3. Tipos de árboles filogenéticos

Existen distintos tipos de árboles filogenéticos. Los **filogramas (phylogram)** miden en las ramas los cambios que ha habido por sitio, por lo que las longitudes de las ramas representan a escala la cantidad de divergencia evolutiva. Tienen la ventaja de mostrar tanto las relaciones evolutivas como la información sobre el tiempo relativo de divergencia de las ramas. Los **cladogramas (cladogram)** muestran la similitud de los distintos elementos, pero las longitudes de sus ramas no son proporcionales al número de cambios evolutivos y, por tanto, no tienen ningún significado filogenético. Los **cronogramas (chronogram)** representan la relación de los elementos de forma temporal.

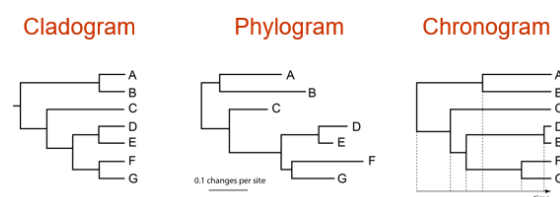


Figura I.5: Tipos de árboles filogenéticos.

## I.4. Inferencia filogenética

Cualquier episodio histórico es, por definición, irrecuperable. La única forma que tenemos de reconstruirlo es a través del estudio de sus efectos. Por ello, la reconstrucción filogenética es un proceso de inferencia: se intenta obtener la mejor estimación posible de una historia evolutiva basada en la información incompleta y con frecuencia ruidosa contenida en los datos. Las evidencias que se emplean se basan en la morfología (comparación entre caracteres de especies), la ultraestructura (cortes vistos al microscopio electrónico), embriología (fases del desarrollo embrionario), la paleontología (registro fósil), la etología (comportamiento animal), la bioquímica y las moléculas.

Un **carácter** es una característica de los taxones que supuestamente es heredada (si no es heredada, no se puede utilizar la filogenia). El **estado de carácter** es el valor específico que toma un carácter en un taxón concreto. Por ejemplo, un carácter sería tener ojos y el estado de ese carácter sería 2 para humanos y 8 para algunas arañas.

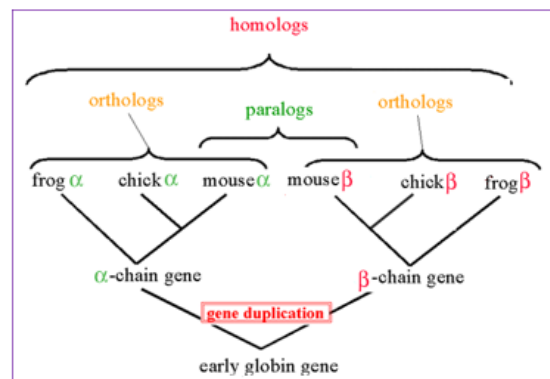
## I.5. Homología

La **homología** es la relación que existe entre dos partes orgánicas diferentes de dos organismos distintos cuando sus determinantes genéticos tienen el mismo origen evolutivo, es decir, cuando un mismo órgano tiene diversas formas y funciones. Los caracteres que se estudian en filogenia deben ser homólogos. Se compara la semejanza de una estructura debido a la herencia común. Por el contrario, la analogía es una estructura semejante a otra o que tiene la misma función, pero cuyo desarrollo embrionario y origen son diferentes. No se presentan en un antepasado común (como en el caso de los caracteres homólogos), si no que es fruto de convergencia evolutiva.

Dentro de la homología se distinguen dos tipos: la ortología y la paralogía. Los **genes ortólogos** son semejantes por pertenecer a dos especies que tienen un antepasado común. Los **genes parálogos** son aquellos que se encuentran en el mismo organismo y cuya semejanza revela que uno procede de la duplicación del otro (y puede adquirir funciones diferentes del gen original). La ortología requiere que se haya producido especiación, mientras que esta no es necesaria en el caso de la paralogía, que puede producirse solo en los individuos de una misma especie. Por ello, idealmente se deben comparar caracteres ortólogos para hacer las reconstrucciones filogenéticas.

### I.5.1. Tipos de homología

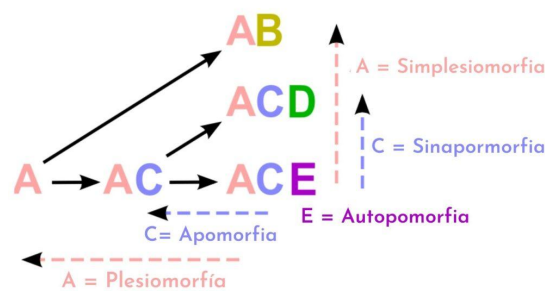
En cladística, se emplean unas clasificaciones de las propiedades de organismos basándose en similitudes derivadas. Se refiere a **plesiomorfia** un estado ancestral o primitivo de un carácter al estar presente en grupos externos y en los ancestros. Cuando un carácter es evolutivamente novedoso pero deriva de otro rasgo perteneciente a un taxón ancestral filogenéticamente próximo, se denomina **apomorfia**. Así, se emplean los adjetivos plesiomórfico y apomórfico en lugar de primitivo y avanzado para evitar juicios de valor sobre la evolución de los caracteres. Una **sinapomorfia** es una apomorfia (carácter exclusivo) compartida por un ancestro común y todos sus



**Figura I.6:** Homología en genes de la hemoglobina.

descendientes. Una **simplesiomorfia** se refiere a una plesiomorfia (carácter ancestral) compartida por dos o más taxa. Finalmente, una **autapomorfia** es un carácter único de un taxón que no aparece en el antepasado.

El prefijo "sin" viene de "compartido". Por tanto, los caracteres sinapomorfos son caracteres apomorfos compartidos, mientras que las simplesiomorfías son plesiomorfías compartidas.



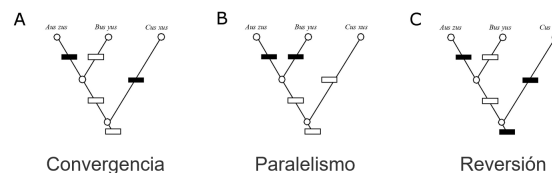
**Figura I.7:** Tipos de homología en el árbol filogenético (tumbado). El carácter A es plesiomórfico al estar en el ancestro. El carácter C es apomórfico al ser una novedad evolutiva. En los nodos terminales, el carácter A se considera simplesiomórfico al estar compartido por los descendientes y ser un carácter ancestral. Por el contrario, el carácter C en los nodos terminales es sinapomórfico por ser un carácter novedoso y estar compartido en el ancestro en el que surgió y sus descendientes. Los caracteres B, D y E son autapomorfos por estar presentes en un único nodo terminal.

## I.5.2. Homoplasia

La homoplasia es el cambio evolutivo paralelo que hace que dos organismos presenten un mismo carácter adquirido independientemente. La **convergencia** se da cuando dos estructuras similares han evolucionado independientemente a partir de estructuras ancestrales distintas y por procesos de desarrollo diferentes. Se considera que el **paralelismo** involucra patrones de desarrollo similares en líneas evolutivas diferentes pero próximas. La diferencia con la convergencia es que en el paralelismo,



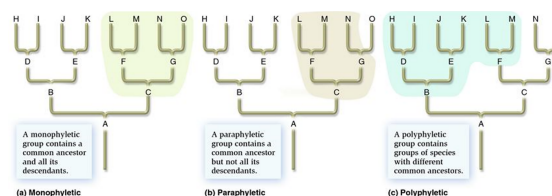
hay un ancestro que no presenta un carácter y dos descendientes directos sí presentan esa novedad evolutiva, mientras que en la convergencia los descendientes con carácter no tienen el mismo ancestro común directo. No obstante, en la práctica, la distinción entre convergencia y paralelismo es un tanto arbitraria porque no existe una regla exacta para limitar la antigüedad del antepasado común. Finalmente, en la **reversión**, un organismo adquiere un carácter de sus antepasados más lejanos. Esto implica que uno o más caracteres adquiridos previamente se han eliminado y se han vuelto a los más anteriores.



**Figura I.8:** *Diferencias entre convergenencia, paralelismo y reversión.*

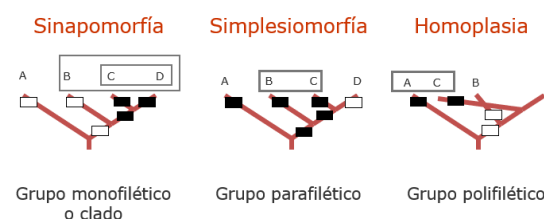
### I.5.3. Agrupamientos

Un **grupo monofilético** es un clado que contiene un ancestro y todos sus descendientes, formando así un solo grupo evolutivo. Un **grupo parafilético** es similar, pero excluye a algunos de los descendientes que han sufrido cambios significativos. Un grupo con miembros de líneas evolutivas separadas se llama **polifilético**, conteniendo así grupos de especies con distintos ancestros comunes.



**Figura I.9:** *Agrupamientos de grupos monofiléticos, parafiléticos y polifiléticos.*

De esa forma, los grupos monofiléticos presentan sinapomorfía, los grupos parafiléticos presentan simplesiomorfía, y los grupos polifiléticos homoplasia.



**Figura I.10:** *Grupos y caracteres que se apoyan.*

### I.5.4. Fenotipo vs moléculas

Tradicionalmente se han empleado los **caracteres fenotípicos** para establecer las relaciones filogenéticas. Esto se debe a que suelen ser caracteres evolutivamente

relevantes y complejos menos proclives a la homoplasia. Además, son los únicos caracteres disponibles en algunos casos como en fósiles o especímenes raros. No obstante, puede haber problemas de codificación de taxones supraespecíficos como terminales (quimares) y se pueden dar casos de subjetividad en la codificación de caracteres. Además, hay un número limitado de caracteres fenotípicos y podemos encontrar taxones altamente autapomórficos.

Recientemente se están empleando **caracteres moleculares** al ser estrictamente heredables y no haber ambigüedades en la codificación. Por ello, determinar el estado de los caracteres es trivial. Hay ciertas regularidades en la evolución de los caracteres moleculares, y éstos son robustos frente a la distancia evolutiva. También son muy abundantes y ofrecen información temporal. El problema de los caracteres moleculares es que son más proclives a la homoplasia al tener solo 4 nucleótidos y 20 aminoácidos. La evolución de estos caracteres es compleja. Además, los árboles de genes no siempre coinciden con los árboles de especies. La determinación de la homología puede ser difícil por duplicación o pérdida de genes y alineamientos.

Se suelen utilizar multitud de genes separados y analizarlos de forma separada. El consenso de análisis separados es una estimación conservadora de la filogenia. Algunos métodos filogenéticos solo se pueden aplicar a ciertos tipos de datos. A nivel de especies, la concatenación de genes diferentes puede ser inapropiada si se da transferencia horizontal de genes, hibridación, duplicación de genes o coalescencia más profunda que el tiempo de divergencia. El conflicto entre caracteres se resuelve teniendo en cuenta toda la evidencia disponible y realizando análisis combinados. Diferentes tipos de datos proporcionan información a diferentes niveles filogenéticos. La señal filogenética aumenta debido a la congruencia entre caracteres de diferentes conjuntos de datos.

Es importante que el conjunto de datos sea lo más completo posible. Es necesario hacer un muestreo de taxones (incluyendo los grupos externos) y genes razonable y justificado.

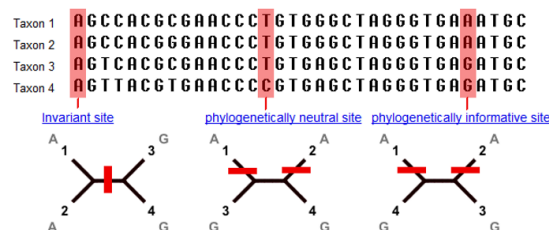
## Capítulo II

# Alineamiento de secuencias

Decidir qué caracteres investigar, y cómo codificarlos, es un primer paso crucial en cualquier análisis filogenético.

## II.1. Tipos de caracteres

Hay **sitios invariables** que no cambian en los distintos taxones. También hay **sitios filogenéticamente neutrales** que son autapomorfías. Los **sitios filogenéticamente informativos** son comunes por pares, por lo que son sinapomorfías.



**Figura II.1:** *Sitios en una secuencia invariantes (no cambian entre los distintos taxones), filogenéticamente neutrales (solo cambia en un taxón) y filogenéticamente informativos (permiten dicotomía).*

Los caracteres pueden ser **binarios 0/1** (presentes o ausentes), **multiestado** o **binarios V/S** (transversiones o transiciones). Los caracteres también pueden ser **discretos** o **continuos**. La codificación de caracteres continuos no se pueden incluir fácilmente en las matrices de caracteres, por lo que se debe realizar una categorización arbitraria. Idealmente, se deben buscar divisiones naturales, es decir, estados discretos de un carácter de variación continua.

## II.2. Ponderación de los caracteres

Se puede emplear un valor relativo de los diferentes caracteres y transformaciones como indicadores de las relaciones filogenéticas entre taxones. Se puede realizar una ponderación uniforme, que minimiza los supuestos del análisis, o una ponderación

diferencial, en la que no todas las características de un organismo tienen el mismo valor como evidencias filogenéticas.

### II.2.1. Ponderación *a priori*

En la ponderación *a priori* de caracteres morfológicos, los taxónomos pueden tener muchas razones para asumir que diferentes caracteres tienen diferente importancia filogenética. Pero eso tiene dos problemas: diferentes opiniones expertas y, en caso de acuerdo, el peso proporcional que se le da a cada carácter. Se introduce precisamente el tipo de subjetividad que el análisis cladístico pretende evitar.

### II.2.2. Ponderación *a posteriori*

El método más utilizado y aplicado, denominado **ponderación implícita**, se basa en Goloboff (1993): la primera vez que un carácter cambia de estado en un árbol, este cambio de estado recibe el peso «1»; los cambios posteriores son menos «costosos» y reciben pesos menores a medida que la tendencia de los caracteres a la homoplasia se hace más evidente. Los árboles que maximizan la función cóncava de homoplasia resuelven el conflicto de caracteres a favor de los caracteres que tienen más homología (menos homoplasia) e implican que el peso medio de los caracteres sea lo más alto posible.

Goloboff reconoce que los árboles con los pesos medios más elevados son los que más «respetan» los datos: un peso medio bajo implica que la mayoría de los caracteres están siendo «ignorados» por los algoritmos de construcción de árboles. Aunque originalmente se propuso con una ponderación severa de  $k=3$ , Goloboff prefiere ahora concavidades más «suaves» (por ejemplo,  $k = 12$ ), que han demostrado ser más eficaces en casos simulados y del mundo real.

$$F = \sum f_i, f_i = \frac{k}{k+(s-m)}$$

m: número mínimo de pasos

s: número de pasos observados

k: constante de concavidad

### II.2.3. Secuencias de ADN

Generalmente, se toma la tasa de sustitución como medida de la fiabilidad de la información filogenética del marcador. Se entiende entonces homoplasia como saturación. Las transversiones evolucionan lentamente y aumentan su frecuencia a medida que pasa el tiempo. Las transiciones se saturan a partir de cierta distancia filogenética, perdiéndose su señal.

El coste de las transformaciones se determina empíricamente mediante matrices de costes (stepmatrices).

## II.3. Polaridad de los caracteres

Necesitamos conocer la polaridad de los caracteres para poder enraizar los árboles. Para ello, se debe establecer qué carácter es ancestral y qué carácter es derivado. Utilizando un **criterio ontogenético**, se ve cómo se forma el carácter durante el desarrollo para poder establecer la polaridad. En caso de que no quede claro tras ese criterio, se compara con el outgroup para establecer el estado primitivo del carácter.

## II.4. Homología de los caracteres moleculares

Cuando analizamos secuencias, asumimos que son de moléculas heredadas de ancestros a descendientes (ortólogos). Cada secuencia está formada por muchos caracteres (cada posición en la secuencia). Por ello, un primer paso es determinar el estado de cada uno de esos caracteres en cada taxón de la matriz. Importante: La homología de los caracteres moleculares, como la de cualquier otro tipo de carácter, es un concepto cualitativo. Las secuencias del gen A de dos taxones son homólogas, o bien no lo son. Igualmente, la posición X en la secuencia de un taxón es homóloga de la posición Y en la secuencia de otro taxón, o bien no lo es. Pero NO puede decirse que las secuencias de dos taxones muestren mayor o menor homología (por ejemplo, en %). Podrán tener diferente porcentaje de similitud (p. ej., % de bases o aminoácidos idénticos en posiciones homólogas), pero o son homólogas o no lo son.

### II.4.1. Aplicación del concepto de homología a los genes: alineamiento de secuencias

Un alineamiento es una hipótesis acerca de la homología posicional de diferentes secuencias de bases o aminoácidos. El alineamiento tiene como objetivo identificar qué posiciones son homólogas en diferentes secuencias. Cada posición de la secuencia (residuo = nucleótido o aminoácido) se interpreta como un carácter que puede tomar diferentes valores (estados de carácter: una de 4 bases, o uno de 20 aminoácidos). El alineamiento asume parsimonia: el cambio evolutivo es improbable, de modo que los segmentos de secuencia coincidentes sirven de guía para identificar posiciones homólogas. Eventualmente se identifican cambios, que cuando son compartidos por varias especies son informativos para la reconstrucción de filogenias.

Las secuencias pueden no tener la misma longitud. Los gaps son marcadores de posición que introducimos en los alineamientos para mantener la homología posicional. Representan eventos de inserción o pérdida denominados indels (del inglés insertion/deletion). La ventaja es que los indels son, en principio, menos propensos a la homoplasia que las sustituciones de bases, muy utilizadas en análisis de parsimonia. No obstante, los indels son difícilmente gestionables por la mayoría de los modelos de evolución molecular.

Es importante elegir un buen alineamiento, ya que la calidad del alineamiento influye en la calidad de la inferencia filogenética.

## II.4.2. Decidir el mejor alineamiento

No existe ningún procedimiento automático para elegir objetivamente el mejor alineamiento: hay que valorar la calidad de los diferentes alineamientos posibles y elegir el que nos parezca mejor. Elegimos como mejor alineamiento el supuesto más razonable de acuerdo con un algoritmo informático y el ojo experimentado. En cualquier caso, es siempre importante examinar el resultado críticamente para valorar si tiene sentido desde un punto de vista biológico.

No todos los alineamientos son igualmente parsimoniosos. Para valorar la calidad de los alineamientos, se han propuesto diferentes mecanismos de puntuación. Se puede realizar una **puntuación por identidad**. Un alineamiento de dos secuencias puede interpretarse como una matriz con dos filas y  $n$  columnas ( $n$  = longitud del alineamiento). Las posiciones (columnas) con idéntico residuo (base o aminoácido) tienen una puntuación = 1. La puntuación del alineamiento es la suma de las puntuaciones de todas sus posiciones. El alineamiento óptimo es el que maximiza la identidad de las columnas. No obstante, los gaps no penalizan, por lo que pueden darse alineamientos con misma puntuación, pero más posiciones de diferencia. Por tanto, se pueden aplicar penalizaciones para los huecos en la secuencia, ya sea introduciendo penalizaciones por la apertura de los huecos o por la extensión de los huecos abiertos. Estos últimos son típicamente menores que las impuestas por apertura. Por ejemplo, se puede aplicar una penalización de -2 por apertura de gap y de -1 por extensión del gap abierto.

ATCG	AT-CG	
ATTG	ATT-G	
1101=3	11001=3	

Izquierda:  $1+1+0+1=3$   
 Derecha:  $1+1-2+1=-1$

$3 > -1 \rightarrow$  el de la izquierda es mejor

**Figura II.2:** Ejemplo de cálculo del mejor alineamiento.

No tiene mucho sentido alinear las secuencias de ADN de los genes codificantes de proteínas. Es mejor traducir las secuencias de ADN a secuencias de aminoácidos y alinear éstas últimas. Existen varios programas para alineamiento múltiple: clustal W/X/Omega, MAFFT, Muscle, T-Coffee, Dialign 2, etc.

## Capítulo III

# Modelos de evolución

Todos los métodos de inferencia y reconstrucción filogenética implican una serie de **supuestos**, aunque éstos no se hagan explícitos:

- Todos los sitios o posiciones cambian independientemente.
- Las tasas de evolución son constantes a lo largo del tiempo y entre linajes.
- La composición de bases es homogénea.
- La verosimilitud de los cambios de base es la misma para todos los sitios y no cambia a lo largo del tiempo.

Esto son asunciones, pero en realidad no son ciertas. Las tasas de evolución no son constantes, las posiciones no cambian independientemente las unas de las otras, la composición de bases no es homogénea (hay mayor porcentaje de GC que de AT) y se pueden dar múltiples cambios en un único sitio que quedan ocultos (si el nucleótido original es C, puede que en un organismo cambie a A y en otro a G). Estos cambios ocultos hacen que las secuencias estén cada vez más saturadas: la mayoría de los sitios que cambian han cambiado antes.

En un contexto filogenético, los modelos predicen el proceso de sustitución de las secuencias a través de las ramas. Describen probabilísticamente el proceso por el que los estados de los caracteres homólogos de las secuencias (posiciones alineadas: nucleótidos o aminoácidos) cambian a lo largo del tiempo.

Los modelos implican por lo general los siguientes **parámetros**:

- Composición: frecuencia de las diferentes bases o aminoácidos.
- Proceso de sustitución: tasa de cambio de uno a otro estado de carácter.
- Otros parámetros (heterogeneidad de tasas): proporción de sitios invariables o agregación de los cambios a lo largo de la secuencia.

### III.1. Modelos frecuentes

El modelo más sencillo es el de Jukes Cantor, el cual asume que todos los cambios son igualmente probables y que la frecuencia de todas las bases es la misma. A partir de este, la complejidad empezó a aumentar, ya que las combinaciones de parámetros son muchas. Algunos de los modelos más frecuentes son:

- Jukes and Cantor (JC69): La frecuencia de todas las bases es la misma (0.25 cada una), y la tasa de cambio de una a otra base es igual.
- Kimura 2-parámetros (K2P): La frecuencia de todas las bases es la misma (0.25 cada una), pero la tasa de sustitución es diferente para transiciones y transversiones.
- Hasegawa-Kishino-Yano (HKY): Como K2P, pero la composición de bases varía libremente.
- General Time Reversible (GTR): La composición de bases varía libremente, y todas las sustituciones posibles pueden tener distintas frecuencias.

Hay programas que ya proponen un modelo a elegir según los datos que se le proporcionen. Cada vez, los modelos son más complejos, y normalmente se utiliza el más complejo.

### III.2. Heterogeneidad de tasas de sustitución

Los modelos anteriores asumen que el cambio es igualmente probable en todas las posiciones de la secuencia y que la tasa de cambio es constante a lo largo de la filogenia. Pero la intensidad de la selección es rara vez uniforme a lo largo de las posiciones, de modo que lo deseable es **modelar la variación de las tasas de sustitución sitio por sitio**.

Para una matriz dada, esperamos observar **posiciones invariables**:

- Porque existen **restricciones funcionales** (selección purificadora relacionada con la función de los genes).
- Porque algunas posiciones **no han tenido ocasión de cambiar**.
- Debido a **homoplasias** que hacen que un sitio aparezca como constante.

La probabilidad de que un sitio sea invariable puede incluirse en los modelos: la verosimilitud de los datos puede aumentar si consideramos que cierta proporción de los sitios son invariables.

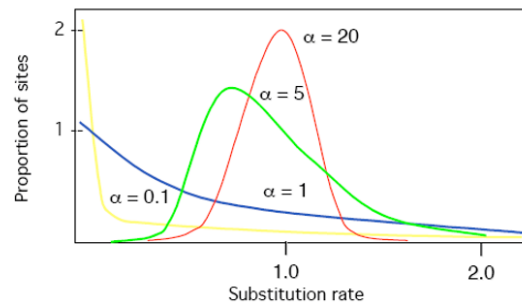
La intensidad de la selección es rara vez uniforme a través de los sitios, de modo que lo deseable es modelar la **variación de las tasas de sustitución sitio por sitio**. Hay dos posibilidades:

- Tasa específica de sitio (posición en el codón, alfa-hélices, etc.)



- Aproximación discreta a una distribución continua (distribución gamma).

La **distribución gamma** se utiliza para modelar la heterogeneidad de las tasas de sustitución entre sitios. La forma de la distribución cambia con diferentes parámetros alfa. Cuanto más bajo es alfa, más concentrados están los cambios en unos pocos sitios. Este parámetro normalmente se calcula por el programa.



**Figura III.1:** Formas de distribuciones gamma con diferentes parámetros alfa.

### III.3. Elegir el mejor modelo

Para obtener el mejor modelo, se suelen utilizar **métodos probabilísticos**. Los mejores valores para cada parámetro son aquellos que, colectivamente, maximizan la **verosimilitud de los datos**. La verosimilitud de un modelo es igual a la probabilidad de los datos (como un alineamiento de secuencias) dada una hipótesis de un modelo de evolución molecular.

Para cada modelo, se calcula la verosimilitud de observar los datos si los cambios de secuencia se producen de acuerdo con el modelo. Los modelos serán tanto más complejos cuanto mayor sea el número de parámetros que contemplen. Se comparan unos modelos con otros mediante tests estadísticos (hLRT) o utilizando los criterios de información de Akaike (AIC) o Bayesiano (BIC).

Al añadir parámetros, el modelo se hace más realista. Sin embargo, cada parámetro es estimado a partir de los datos: cuantos más parámetros añadamos, mayor será la varianza de nuestras estimaciones. Por ello, un modelo demasiado “realista” (complejo) puede provocar un excesivo término error, perdiendo potencia estadística. Los criterios de información tienen en cuenta ambas características de los modelos:

- El ajuste del modelo (es mejor el modelo que hace más verosímil observar esos datos).
- La complejidad del modelo (de dos modelos igualmente verosímiles, el más simple es mejor).

#### III.3.1. Selección de esquemas de partición

Elegir la mejor forma de particionar los datos es casi tan importante como elegir el modelo de sustitución. Distintas partes del alineamiento (posiciones de codón,

codificantes frente a no codificantes, etc) pueden ajustarse mejor a diferentes modelos de sustitución o diferentes parámetros de un mismo modelo. Algunas particiones con parámetros de modelos similares pueden combinarse de modo que el número de parámetros final se reduce. La elección del mejor esquema de partición se hace también utilizando criterios de información (AIC y/o BIC). Hay métodos de búsqueda simultánea de los mejores modelos de sustitución y esquemas de partición, como [PartitionFinder](#).

## Capítulo IV

# Métodos filogenéticos de inferencia

Los pasos generales en el proceso de reconstrucción filogenética son:

1. Diseño experimental: selección del ingroup y outgroup, selección de los marcadores moleculares
2. Recolección de datos de secuencia homóloga
3. Ensamblaje de la matriz de secuencias
4. Alineamiento de la secuencia
5. Selección del modelo
6. Inferencia filogenética
7. Construcción del árbol filogenético: soporte estadístico, testar hipótesis filogenética, estimación del tiempo de divergencia

### IV.1. Búsqueda de árboles

Sólo hay una manera de construir el primer árbol sin raíz, uno con tres puntas (nodos terminales) y tres ramas. Cada vez que añadimos un taxón, se crean dos ramas. Un árbol con  $n$  puntas (taxones) tendrá por tanto  $2n-3$  ramas.

A partir de unas pocas especies, las búsquedas de árboles sin raíz son exhaustivas y computacionalmente demasiado exigentes. Por ello, se realiza una **búsqueda heurística**:

1. Construir el árbol inicial (Ej., mediante adición secuencial de taxones) y determinar su longitud (ver el número de cambios de un taxón a otro).
2. Construir un conjunto de "árboles vecinos" o alternativos haciendo pequeñas reordenaciones en el árbol inicial, y determinar las longitudes de cada nuevo árbol.

3. Si cualquiera de los árboles vecinos es mejor que el inicial (tienen un menor número de pasos o cambios evolutivos, es decir, la hipótesis se ajusta mejor a los datos): retenerlo y usarlo como punto de partida para una nueva ronda de reordenaciones (es posible que varios de estos árboles sean igual de buenos).
4. Repetir pasos 2 y 3 hasta encontrar un árbol que es mejor que todos sus vecinos.
5. Este árbol es un óptimo local (¡no necesariamente un óptimo global!)

El procedimiento semeja un paseo en un paisaje montañoso, donde nos interesa alcanzar la cumbre más alta (hill climbing).

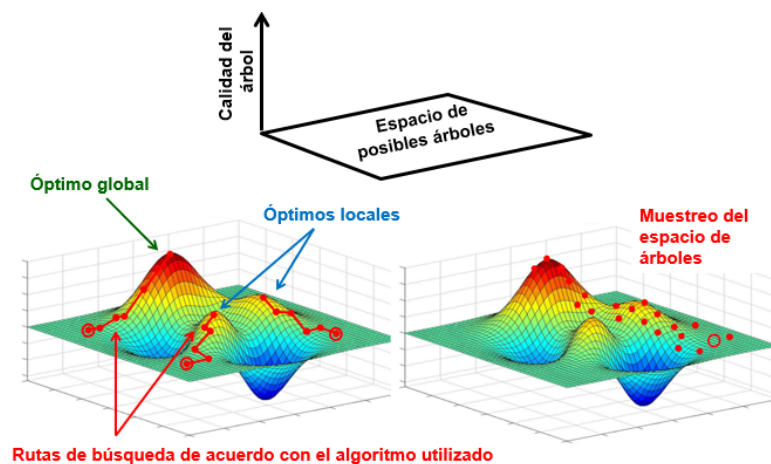


Figura IV.1: Búsqueda heurística del mejor árbol.

Hay varios algoritmos de reordenación de ramas (branch swapping):

- **Nearest Neighbour Interchange (NNI)**: intercambia dos vecinos por cada rama interna.
- **Subtree Pruning and Regrafting (SPR)**: se corta un clado (subárbol) y se empalma en todas las ramas del resto del árbol, usando el punto de corte del subárbol como punto de unión. Realmente, NNI es un subconjunto de SPR.
- **Tree Bisection and Reconnection (TBR)**: se divide el árbol en dos partes y se reconectan los subárboles usando todos los posibles pares de ramas. NNI y SPR son subsets de TBR.

El espacio de árboles puede estar poblado por mínimos locales e islas de árboles óptimos.

## IV.2. Árboles de consenso

A menudo existen varios árboles candidatos a ser el cladograma más parsimonioso<sup>1</sup>. En el siguiente ejemplo, hay tres cladogramas diferentes que son igualmente

<sup>1</sup>véase explicación de parsimonia en el tema 5

## Branch swapping

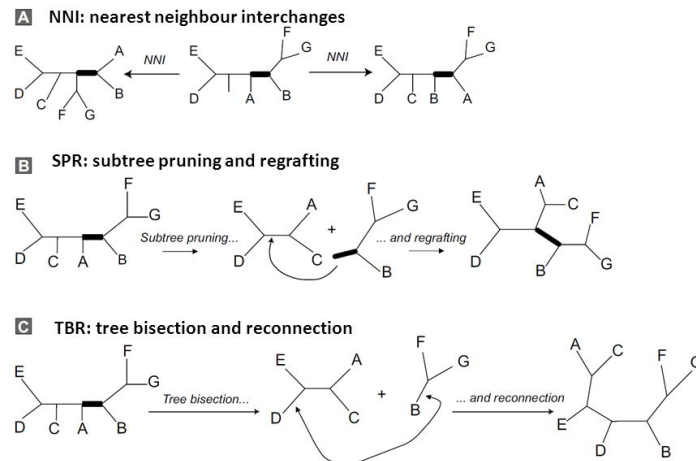


Figura IV.2: Esquemas de los distintos algoritmos de reordenación de ramas.

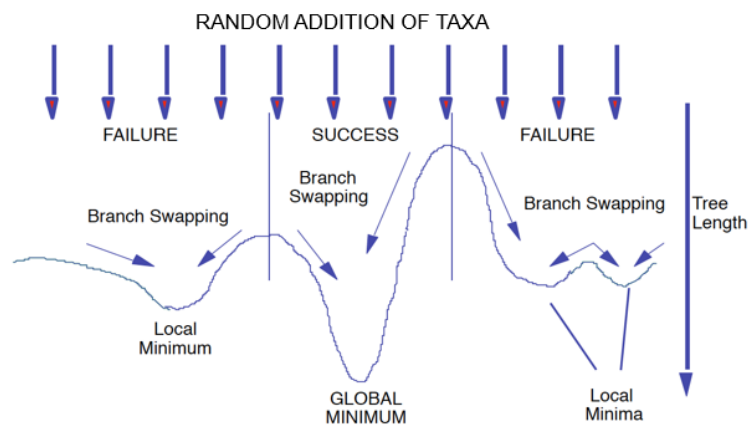
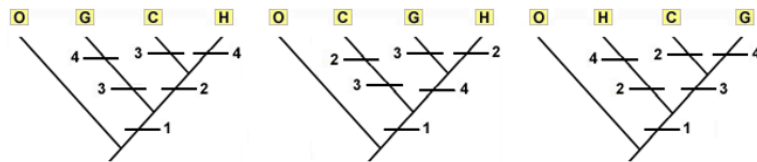


Figura IV.3: Posibles resultados de árboles tras la adición de taxones.

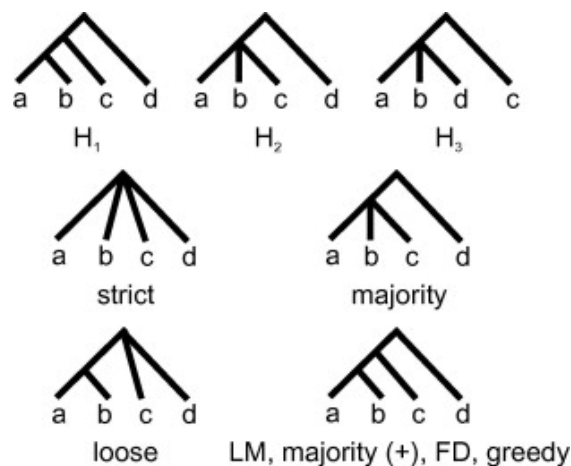
parsimoniosos para los 4 caracteres estudiados (1-4) en cuatro especies de homínidos. Si aumentamos el número de caracteres y el de taxones, la cantidad de cladogramas igualmente parsimoniosos se dispara y se hace inmanejable. Por lo tanto, es conveniente contar con formas estandarizadas de resumir los puntos de acuerdo entre cladogramas rivales para llegar a formar un “árbol de consenso”.



**Figura IV.4:** Tres cladogramas igualmente parsimoniosos de cuatro especies de homínidos: orangutanes, gorilas, chimpancés y humanos.

Un árbol de consenso no es otra cosa que un árbol que combina los agrupamientos preferidos a partir de los cladogramas rivales de un determinado grupo de taxones, de tal forma que los agrupamientos discutibles (contenciosos, ambiguos) se condensan en puntos con múltiples ramas (politomías= nodos que portan múltiples ramas). Existen diferentes formas para construir árboles de consenso, pero los tres métodos más comunes son:

- **Árbol de consenso estricto:** conserva sólo los agrupamientos que comparten todos los cladogramas rivales.
- **Árbol de consenso semi-estricto o loose:** conserva todos los agrupamientos que no son contradictorios en los cladogramas rivales.
- **Árbol de consenso de regla de la mayoría:** conserva todos los agrupamientos que son apoyados por la mayoría de cladogramas rivales, aunque contradigan a la minoría. Este es el que más se utiliza.



**Figura IV.5:** Distintos tipos de árboles de consenso.

Cuando tenemos muchos taxones entre los cuales hay mucha homoplasia, los árboles de consenso estricto van a mostrar gran número de politomías que no pueden resolverse ya que existen conflictos de agrupamiento entre los cladogramas rivales.

En esos caso, los árboles de consenso basados en la regla de la mayoría, ofrecen al menos bastantes más hipótesis de trabajo para investigar las relaciones filogenéticas. El ejemplo está basado en más de 70 especies de erizos irregulares; para mayor claridad se han omitido los nombres de las especies. Se estimó que había más de 70.000 cladogramas igualmente parsimoniosos, lo que refleja un gran componente de homoplasia entre los caracteres analizados. El árbol de consenso estricto resultante, ofrece por tanto poca resolución, y la mayoría de los taxones de agrupan en una gran politomía. Por el contrario, el árbol de consenso de regla de la mayoría, resuelve un número de clados mucho mayor que podrán comprobarse en sucesivos análisis. Los valores numéricos para el árbol de consenso de regla de la mayoría miden el porcentaje de cladogramas igualmente parsimoniosos que apoyan cada grupo resuelto.

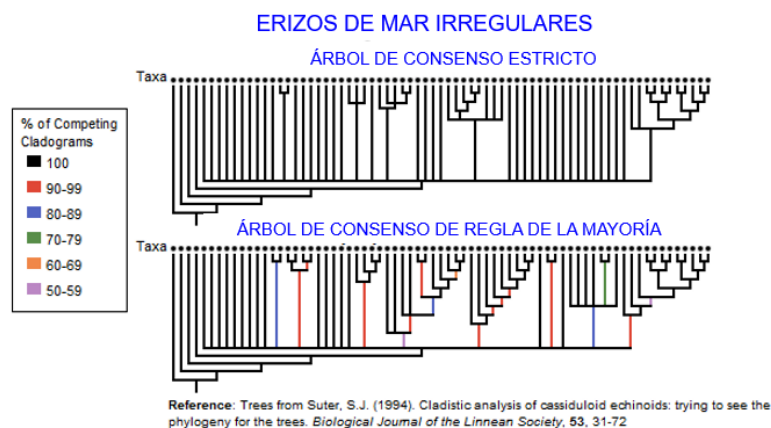


Figura IV.6: Árboles filogenéticos de erizos de mar irregulares.

### IV.3. Medidas de soporte: confianza en el árbol

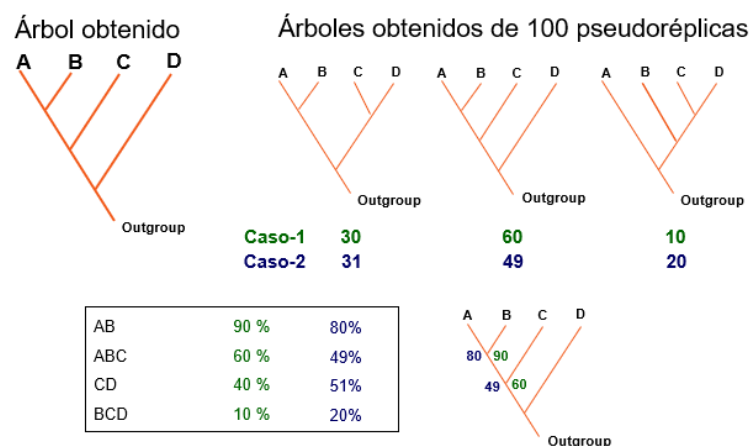
La mayor parte de las medidas científicas van acompañadas de una estima de su precisión del tipo. En el caso de la inferencia filogenética, no es suficiente con estimar hipótesis filogenéticas, sino que es necesario indicar una estima de la confianza que dicha hipótesis presenta. Esto se debe a:

- **Error de muestreo:** Nuestros análisis están basados en muestras, pero sólo tenemos una muestra de los datos. Los valores estimados a partir de muestras de una población raramente van a coincidir con el valor real. Una forma de calcular este error es tomando múltiples muestras y comparando las estimas obtenidas entre sí.
- **Error sistemático:** Asociado a la metodología y las asunciones de los análisis.

Las formas de dar apoyo y soporte a un nodo se pueden dar de forma cualitativa (soporte de Bremer), remuestreo (bootstrapping, jackknife) y probabilístico (probabilidad posterior bayesiana):

- **Soporte de Bremer o Decay Index:** se calcula la diferencia en el número de pasos entre el árbol óptimo y el mejor árbol en el que no aparece el clado en cuestión.

- Remuestreo por bootstrapping:** se remuestran los caracteres al azar con reemplazamiento múltiples veces. Se realiza el análisis con cada nueva pseudoréplica utilizando los mismos parámetros que en el análisis original. Se analiza la coincidencia entre las topologías obtenidas resumiéndolas en un majority-rule consensus tree. Las pseudoréplicas se construyen a partir de la matriz original con reemplazamiento para construir una nueva matriz del mismo tamaño que la original. La frecuencia con que aparece un determinado grupo es una medida de la estabilidad de ese grupo. Estos valores se muestran en un árbol de majority-rule consensus y se da información adicional en una tabla (de biparticiones). Los valores de bootstrap son conservadores; son un índice relativo del soporte estadístico de los grupos, proporcionado por los datos que se están analizando bajo un método de análisis concreto: valores altos de bootstrap nos indican la existencia de una señal filogenética “fuerte” en los datos. Estudios realizados con datos empíricos y simulaciones han indicado que un 70 % de soporte de bootstrap puede considerarse apoyo razonable para una relación determinada. No obstante, este número se puede modificar según se necesite.
- Remuestreo por Jackknife:** Jackknife es muy similar al bootstrap, sólo se diferencia en la estrategia de remuestreo de los caracteres. Una cierta proporción de los caracteres es eliminada al azar (por ej.: 50 %). No hay reemplazo, por lo que la matriz es más pequeña. Se analizan las pseudoréplicas y los resultados se resumen en un majority-rule consensus tree. Jackknifing y bootstrapping suelen dar resultados similares y se interpretan de forma similar. Jackknife se está utilizando cada vez menos y se está reemplazando por bootstrap al estar reduciendo los datos.



**Figura IV.7:** Ejemplos del cálculo de bootstrap. En el primer caso (mostrado en verde), tras 100 pseudoréplicas, han salido tres árboles con frecuencias de 30, 60 y 10. Tanto en el primer como en el segundo árbol, los taxones A y B se han relacionado juntos, por lo que esa dicotomía tiene un soporte de bootstrap de 90 (60 + 30). La siguiente relación más soportada, con un bootstrap de 60, es la de relacionar el taxón C con el antepasado común de A y B, por lo que el árbol final muestra esa variante (la otra opción sería relacionar C con D, como hacen los otros dos árboles, pero su frecuencia es de 40). En el segundo caso (mostrado en azul), las frecuencias han cambiado. Ahora, la relación de A y B pasa a tener un soporte de 80 (31 + 49), y así sucesivamente.



## Capítulo V

# Máxima parsimonia - Cladística

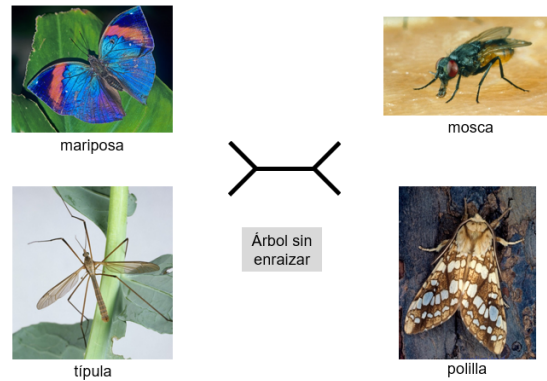
La cladística es un método de análisis de la sistemática filogenética que busca reconstruir las “genealogías” de los organismos y elaborar clasificaciones que las reflejen. Descansa sobre el axioma fundamental de que en la naturaleza, como resultado de la evolución, existe un orden que se manifiesta en las similitudes de los caracteres. Determina las relaciones evolutivas entre los organismos basándose en los caracteres relativamente derivados o apomórficos (novedades evolutivas). La reconstrucción filogenética consiste en identificar todos los grupos monofiléticos que existen en una muestra de taxones, que son aquellos definidos por sinapomorfías (caracteres derivados compartidos).

### V.1. Construir el árbol

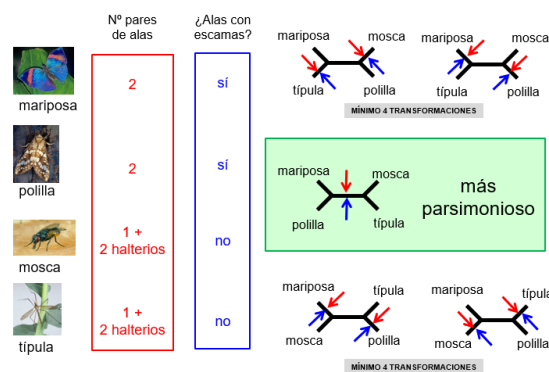
Desconocemos el aspecto del antecesor común más reciente de las especies y el modo en que están emparentadas, por lo que comenzamos a analizar sus relaciones buscando las diferentes formas en que pueden ser conectadas. Este ejemplo implica mayor similitud entre red (network) o árbol sin enraizar con ramificación dicotómica que conecta un grupo de taxones. No tiene raíz que conecte con un antecesor común; es como un mapa filogenético visto desde arriba, con el antecesor común oculto por sus descendientes. Es de ramificación dicotómica porque sólo tres ramas se juntan en cada unión o nodo; cada línea se divide siempre en dos ramas. Se pueden obtener otra red cambiando la posición de las especies. Existen tres posibles formas de unir 4 especies en una red de ramificación dicotómica.

La selección de la explicación más sencilla de la distribución de los diferentes estados de caracteres se denomina **principio de parsimonia**. Así, se revisan los caracteres y se construye la matriz para ver cómo se sitúan los estados de los caracteres en las redes posibles. Aplicando el principio de parsimonia, se rechazan aquellos árboles que tengan más transformaciones en la red y se queda el más parsimonioso, es decir, el que necesita menos transformaciones evolutivas para explicar la distribución de los estados de los caracteres analizados.

Ante distintas hipótesis sobre las relaciones filogenéticas de los organismos dados, elegimos la más parsimoniosa, esto es, la que tenga una longitud menor (menor número de cambios), un mayor número de homologías y un menor número de homoplasias. Si



**Figura V.1:** Ejemplo de un árbol sin enraizar entre cuatro especies. Existen dos alternativas más árboles sin enraizar: mariposa-polilla con mosca-tipula y mariposa-mosca con tipula-polilla.

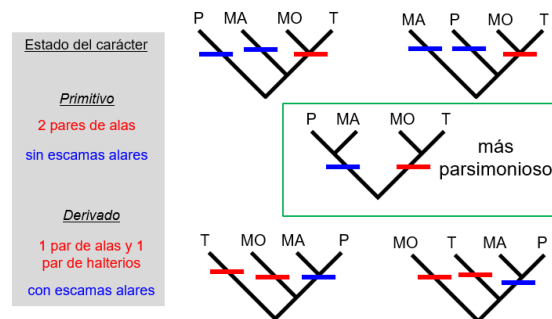


**Figura V.2:** Matriz de los estados de caracteres del ejemplo anterior. Tanto la mariposa como la polilla tienen dos pares de alas y las alas tienen escamas, mientras que la mosca y la típula tienen un par de alas y dos halterios y las alas no tienen escamas. A partir de ahí, se calculan las transformaciones necesarias para explicar la distribución del árbol y se elige el más parsimonioso, es decir, el que necesite un menor número de cambios.

hay varios árboles o hipótesis igualmente parsimoniosos, no podremos elegir entre ellos. La longitud de un árbol indica el número de transformaciones evolutivas necesarias para explicar los datos dada una topología de árbol concreta. Corresponde al número de cambios de estado que se producen en el árbol.

## V.2. Enraizar el árbol

El siguiente paso es enraizar el árbol. La raíz nos permite determinar el lugar donde se encuentra el ancestro común. Para ello, se debe elegir la polaridad de los caracteres, es decir, conocer en qué orden se produjeron las transformaciones. Conocer qué estado del carácter es primitivo y cuál es derivado nos ayudará a elegir el cladograma con la distribución más sencilla de estados derivados (el más parsimonioso).

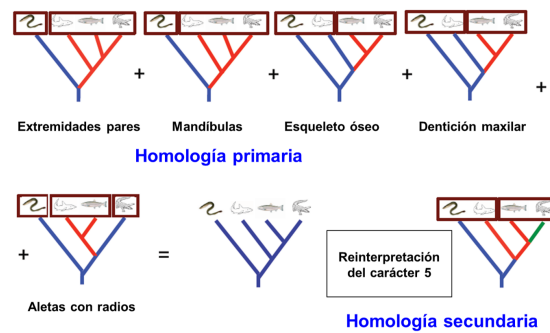


**Figura V.3:** Opciones de árboles enraizados siguiendo el ejemplo anterior. Primero, sería importante conocer en qué orden se produjeron las transformaciones, si derivaron los halterios del segundo par de alas o si fue a la inversa. Supongamos que hay evidencias que sugieren que los halterios derivaron del segundo par de alas (que era el estado primitivo) y que la adquisición de escamas en las alas es otro estado de carácter derivado. En los posibles árboles, marcamos el lugar donde debe ocurrir la transformación de cada carácter de primitivo a derivado. El árbol más parsimonioso es aquel que necesita menos transformaciones para explicar la distribución de los estados derivados de los caracteres, y el que probablemente mejor registre las relaciones evolutivas de las cuatro especies de insectos.

La máxima parsimonia asume los siguientes principios:

- Principio auxiliar de Hennig: Ante caracteres similares y en ausencia de evidencia que indique evolución paralela o convergencia, siempre se asume que dichos caracteres son homólogos.
- Regla de agrupación de Hennig: Las sinapomorfías son evidencias de relaciones de ancestro común, mientras que las simplesiomorfías, las convergencias y los paralelismos no proporcionan evidencias sobre el ancestro común.

Una auténtica homología debe circunscribir un grupo consistente con los especificados por otras homologías. Para comprobar esto, se realiza un **test de congruencia**. Así, se considera homología primaria cuando coincide, mientras que cuando se debe reinterpretar un carácter, se considera homología secundaria.



**Figura V.4:** Árboles filogenéticos entre anguilas, tiburones, peces y cocodrilos basados en los caracteres de extremidades pares, mandíbulas, esqueleto óseo, dentición maxilar y aletas con radios. La mayoría de los árboles muestran una distribución similar de los taxones, mostrando así homología primaria. No obstante, el carácter aletas con radios es diferente, al relacionar dos taxones que en los demás caracteres no estaban relacionados. Ese carácter se debe reinterpretar, siendo así homología secundaria.

En el cladismo, se busca maximizar la congruencia entre caracteres, y así minimizar la incongruencia (homoplasia). En una aplicación computacional de la máxima parsimonia, se sigue un criterio de optimización (criterio para escoger entre diferentes cladogramas / relaciones filogenéticas), en el cual el cladograma preferido es aquél que tiene el menor número de transformaciones entre estados de carácter (pasos). Determinar los estados ancestrales implica poder identificar cambios de estado de carácter.

### V.3. Valoración de árboles

Encontrar el cladograma más parsimonioso es la meta final de los análisis filogenéticos, aunque preguntas sobre su fiabilidad pueden permanecer. Para comparar objetivamente diferentes cladogramas y valorar el ajuste de los caracteres se han desarrollado medidas para cuantificar el alcance del problema de la homoplasia en los cladogramas obtenidos. Se comentan tres medidas:

- Longitud del árbol: mide el número mínimo de transformaciones que se requieren en un determinado cladograma a partir de la matriz de caracteres (los árboles de longitud mínima para una matriz se llaman árboles de Wagner).
- Índice de consistencia (CI): muestra la proporción de transformaciones que no se repiten (es decir, que no son homoplasias) en un cladograma.
- Índice de retención (RI): estima el alcance de las sinapomorfías entre los estados de los caracteres en un cladograma determinado.
- Índice de retención re-escalado (RCI): combina los dos anteriores, situando la homoplasia observada en una escala desde la mínima (0) a la máxima posible (1).