

Table of Contents

1. Introducción a la filogenética
 1. Estudios filogenéticos
 2. Árboles filogenéticos
 1. Partes de un árbol filogenético
 2. Representación de los árboles
 3. Tipos de árboles
 4. Agrupamientos
 5. Politomías
 3. Conceptos básicos
 1. Homología
 2. Homoplásia
 3. Fenotipo vs moléculas
 2. Alineamiento de secuencias
 3. Modelos de evolución molecular
 4. Métodos filogenéticos de inferencia
 5. Máxima parsimonia
 6. Máxima verosimilitud
 7. Inferencia Bayesiana
 8. Cronogramas - árboles temporales
 9. What do we need to build a phylogenetic tree? - Summary
10. Ejercicios

Introducción a la filogenética

Estudios filogenéticos:

La filogenética puede ser estudiada de diversas maneras.

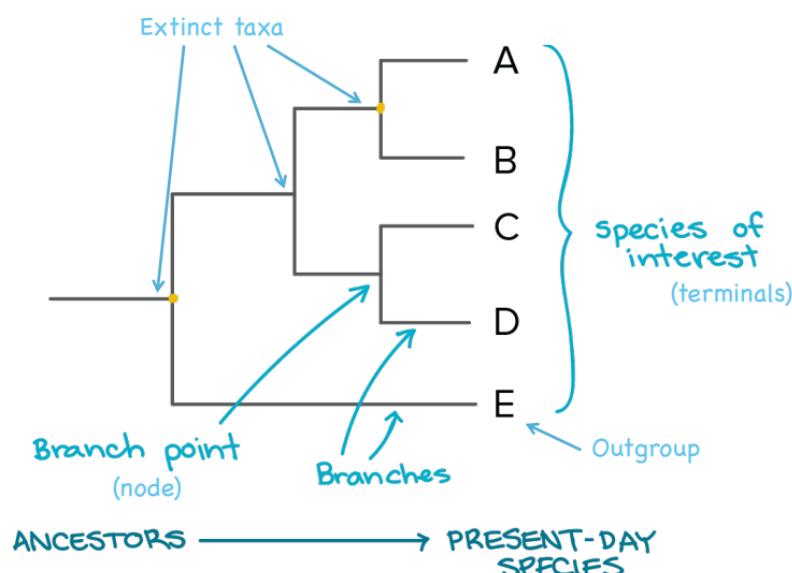
- Registros fósiles:
 - PROS: contienen información sobre la morfología de los antepasados de las especies actuales y la cronología de sus divergencias. Esto permite datar las filogenias.
 - CONTRAS: utilizar registros fósiles para determinar relaciones filogenéticas puede producir **sesgos** porque:
 - 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
 - las descripciones de los rasgos morfológicos son a menudo ambiguas (múltiples factores genéticos).
- Datos moleculares: en la forma de secuencias de ADN o de proteínas. Debido a que los genes son el medio para registrar las mutaciones acumuladas, éstos pueden servir como "fósiles moleculares".

- **PROS:** 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.

Arboles filogenéticos:

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.

Partes de un árbol filogenético:



Nodos externos o terminales.

- Se denominan **grupos hermanos** a los nodos terminales que parten de un mismo nodo interno, es decir, dos taxones que comparten 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 arbitrariamente para poder colocar la raíz donde se estima correcto.
- Todas las especies que se desarrollan desde una rama de la raíz se denominan **grupo interno o ingroup.**

Nodos internos:

Son hipótesis evolutivas de posibles ancestros comunes de los cuales normalmente faltan datos para confirmar o descartar la teoría.

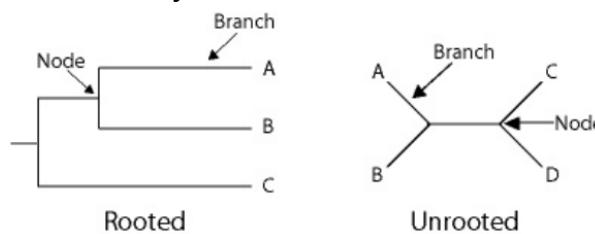
Ramas (branches) que unen los nodos.

En las distintas ramas se pueden representar la transformación de caracteres que aparecen a nivel genético y que se transmiten por herencia.

Raiz.

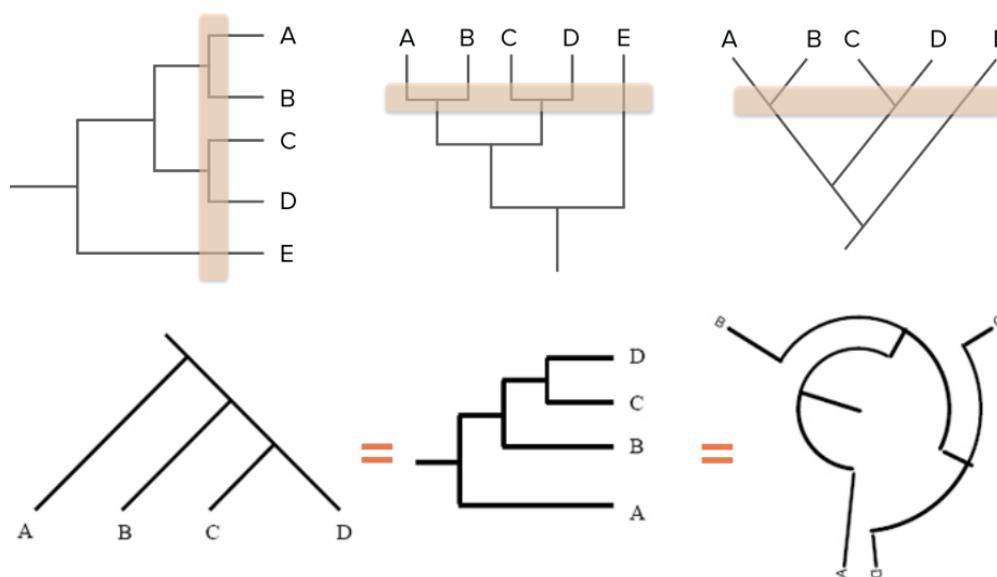
Los árboles filogenéticos se pueden representar sin enraizar o enraizado.

- **Sin raíz:** Un árbol filogenético que no asume conocimiento de un ancestro común, solo posiciones de los taxones para mostrar sus relaciones relativas (no hay dirección de un camino evolutivo).
- **Con raíz:** Para describir la dirección de la evolución se necesita un árbol filogenético donde todas las secuencias bajo estudio tienen un ancestro o nodo raíz común (*más informativo*).



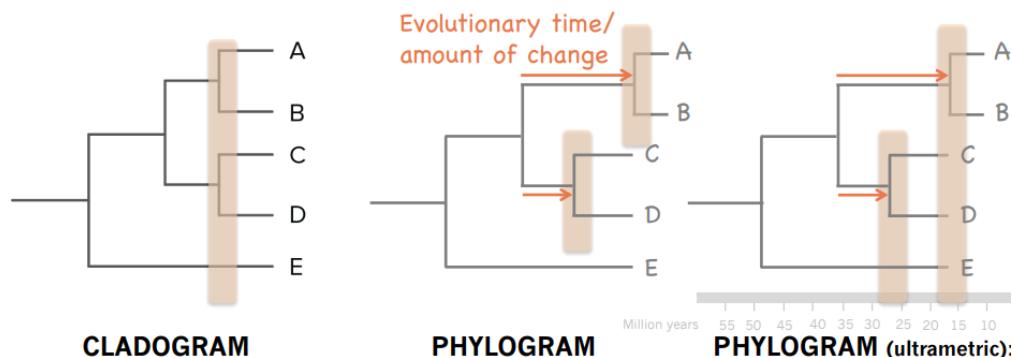
Representación de los arboles:

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.



Tipos de arboles:

1. **Cladogramas (CLADOGRAM)** used in cladistics BUT a cladogram is not an evolutionary hypothesis since it does not contain information about how ancestors and descendants are related or how much descendants have changed through time. They are a SIMPLISTIC REPRESENTATION (OK for morphological matrices).
2. **Filogramas (PHYLOGRAM):** a branching tree where the branch lengths indicate the amount of **evolutionary change** inferred from the analysis. Tienen la ventaja de mostrar tanto las relaciones evolutivas como la información sobre el tiempo relativo de divergencia de las ramas.
3. **PHYLOGRAM (ultrametric) or CRONOGRAM:** ages assigned to each node using molecular clocks. Representan la relación de los elementos de forma temporal.



[phyl(o) gr. 'raza', 'estirpe']

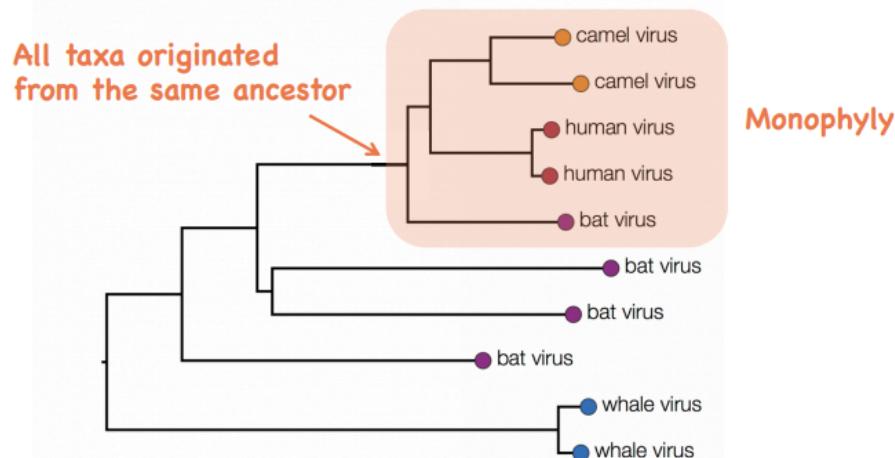
[klad(o) gr. 'rama']

+ [-gram-ma gr. 'representación gráfica']

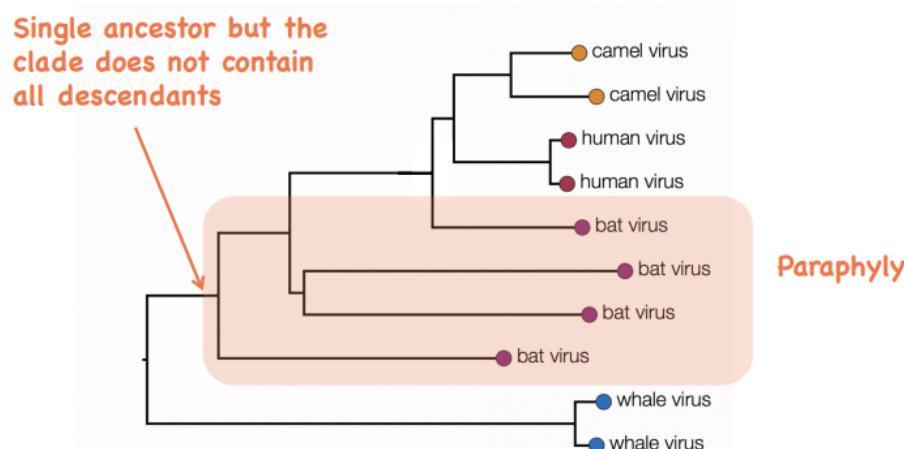
[khron(o) gr. 'tiempo']

Anatomy of the phylogenetic tree / Agrupamientos:

- **Monophyly / grupo monofilético:** (los que nos interesan en filogenia) Un conjunto de taxones que comparten una sinapomorfía, es decir, que contiene un ancestro y todos sus descendientes, formando así un solo grupo evolutivo = un clado.

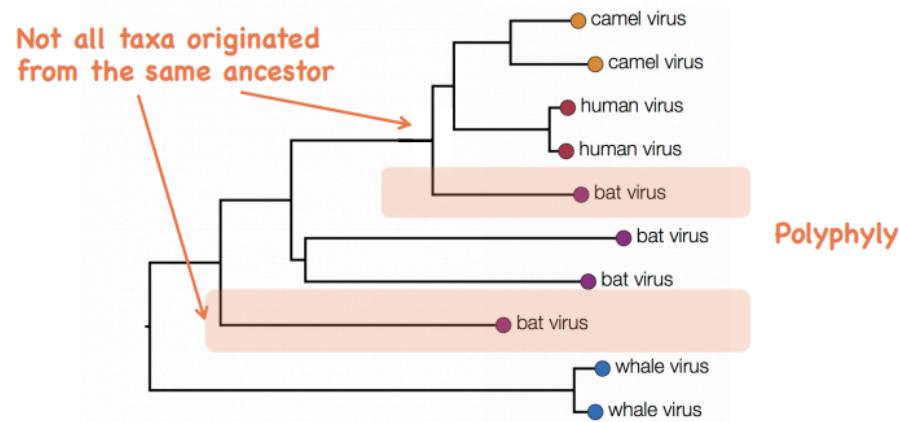


- **Paraphyly / grupo parafilético:** Un conjunto de taxones con una simplesiomorfía en común, incluye el antecesor común más reciente y algunos de sus descendientes, pero no a todos.



- **Polyphyly / grupo polifilético:** Un grupo de taxones agrupados por la presencia de un carácter homoplásico. Es un grupo en el que el antecesor común más reciente, y quizá también algunos de sus

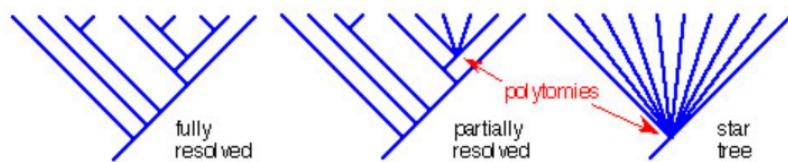
descendientes, no son miembros de este grupo, es decir, un grupo con miembros de líneas evolutivas separadas.



Grupos y caracteres que se apoyan

Politomías:

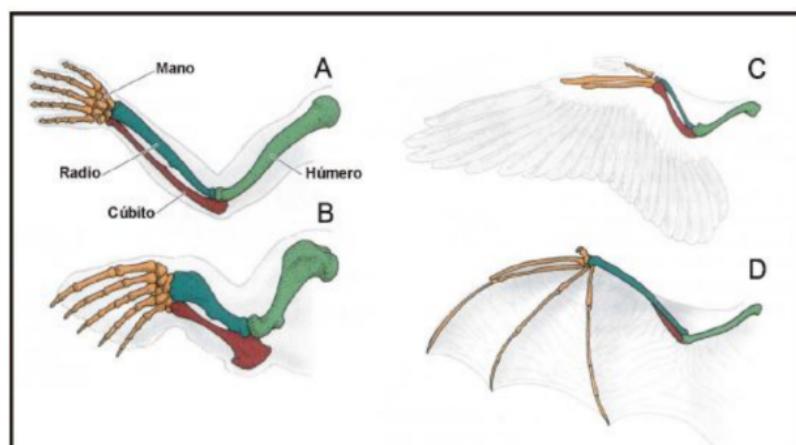
- **Dicotomía:** Cuando todas las ramas en un árbol filogenético se bifurcan. Los árboles filogenéticos se consideran resueltos cuando sus ramas se distribuyen dicotómicamente.
- **Politomía:** si de un nodo surgen más de dos ramas (descendientes). Los árboles no resueltos presentan politomías.



Conceptos básicos:

Homología

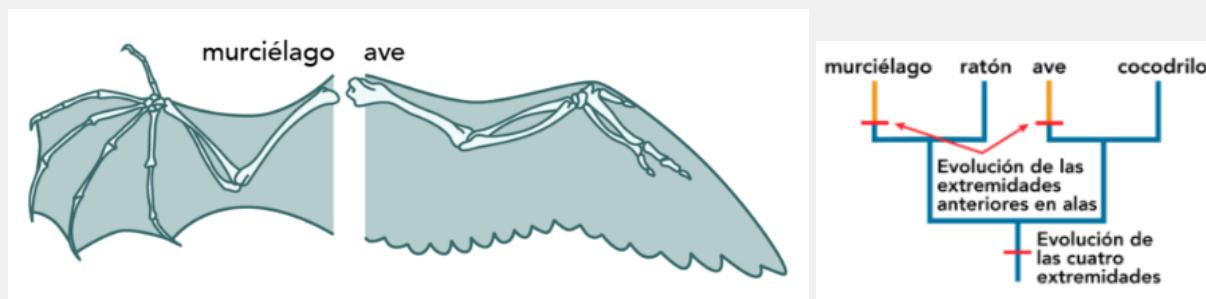
La homología es la relación que existe entre dos partes orgánicas diferentes de dos organismos distintos cuando sus determinantes genéticos tienen el mismo origen evolutivo, es decir, cuando un mismo órgano tiene diversas formas y funciones.



El mismo órgano diversas formas y funciones. Semejanza en la estructura debido a la herencia común.

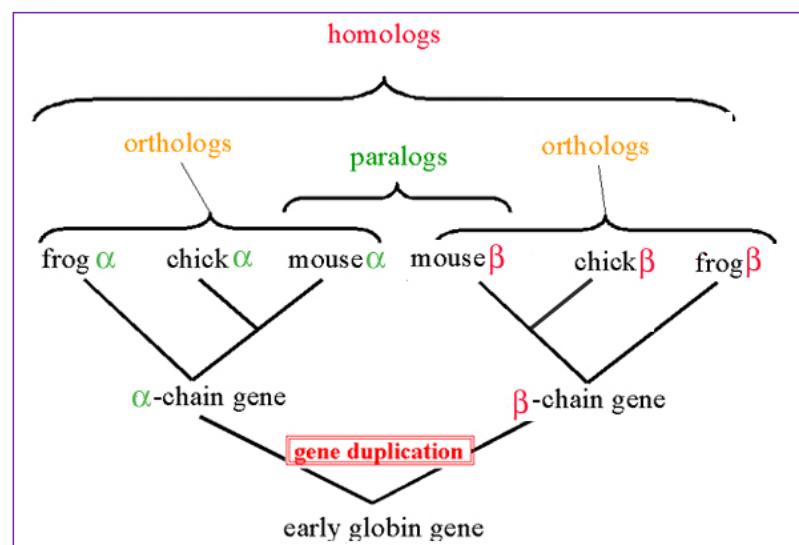
Los caracteres que se estudian en filogenia deben ser **homólogos**. Se compara la semejanza de una estructura debido a la herencia común.

Las alas de las aves y las de los murciélagos son **análogas** — es decir, sus orígenes evolutivos son independientes, pero se parecen superficialmente porque evolucionaron para realizar la misma función. Las analogías son el resultado de la **evolución convergente** o **paralelismo**.



Por lo tanto, es interesante que mientras las alas de las aves y de los murciélagos son análogas como alas, como miembros anteriores son homólogos. Las aves y los murciélagos no heredaron las alas de un antepasado común alado, pero sí heredaron las extremidades anteriores de un antepasado común con extremidades anteriores.

En genética y biología molecular, también existe **homología en las secuencias**. Los caracteres o estados son homólogos si derivan del mismo carácter en el ancestro común más cercano. Se distinguen dos tipos: la **ortología** y la **paralogía**.



A gene that has diverged as a result of a speciation event is called an ortholog. Orthologs will generally retain the same function after the speciation event—this is how ‘transfer of annotation’ is possible. But they may not have the same name. If two genes diverge as a result of a gene duplication event, they are called paralogs. Generally, paralogs will take on a different-but-related gene function, while their cousins—the orthologs—will retain the same function through the course of evolution.

Genes Ortólogos

Son semejantes por pertenecer a dos especies que tienen un antepasado común. Requiere que se haya producido especiación. En la imagen son el conjunto de hemoglobinas beta y alfa de los grupos.

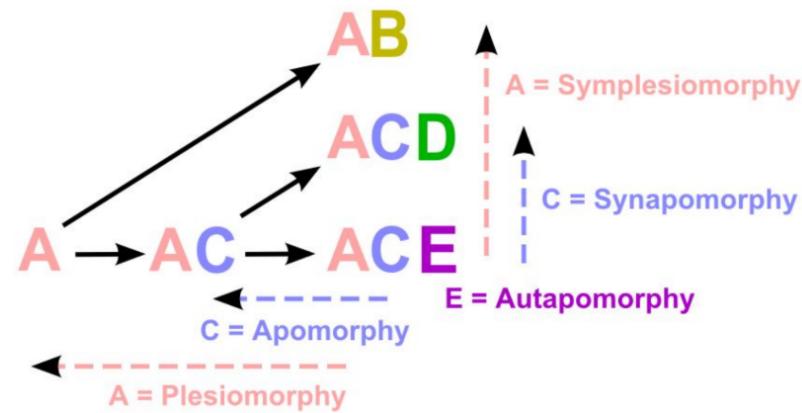
Genes Parálogos

Son aquellos que se encuentran en el mismo organismo y cuya semejanza revela que uno procede de la duplicación del otro (y si hay posterior mutación, puede adquirir funciones diferentes del gen original). La especiación no es necesaria, la paralogía puede producirse solo en los individuos de una misma especie.

Note: Idealmente se deben comparar caracteres **ortólogos** para hacer las reconstrucciones filogenéticas.

Tipos de homología:

Clasificaciones de las propiedades de organismos basándose en similitudes derivadas.



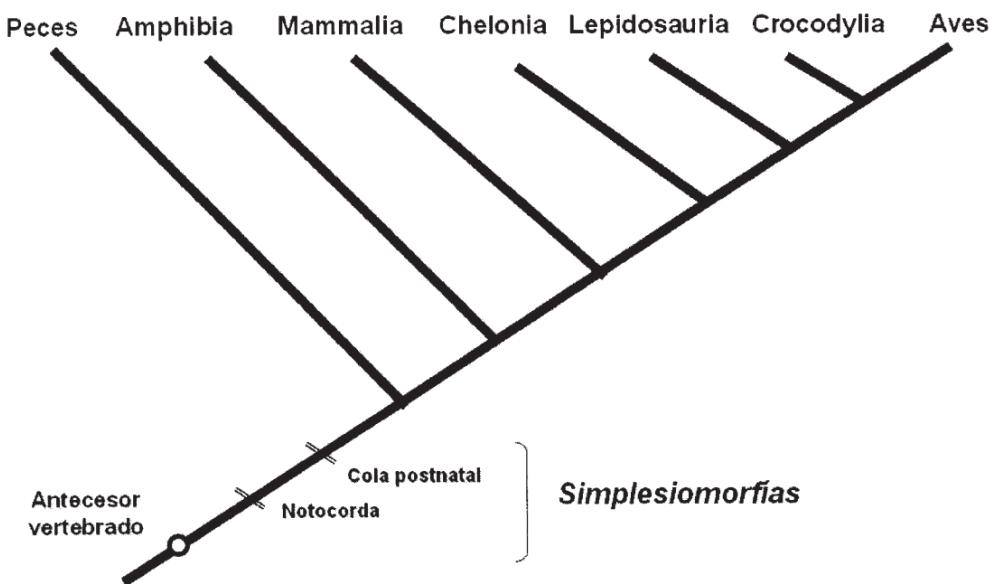
Tipos de homología en el árbol filogenético. El carácter A es plesiomórfico al estar en el ancestro. El carácter C es apomórfico al ser una novedad evolutiva. En los nodos terminales, el carácter A se considera simplesiomórfico al estar compartido por los descendientes y ser un carácter ancestral. Por 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 autopomorfos por estar presentes en un único nodo terminal.

Se distinguen dos tipos de estados en los caracteres homólogos:

Plesiomorfía (ancestral character state) se refiere al estado ancestral (o primitivo) de un carácter que comparten distintas especies por heredarlo del antepasado común; en el ejemplo se presenta en los ancestros y los grupos externos.

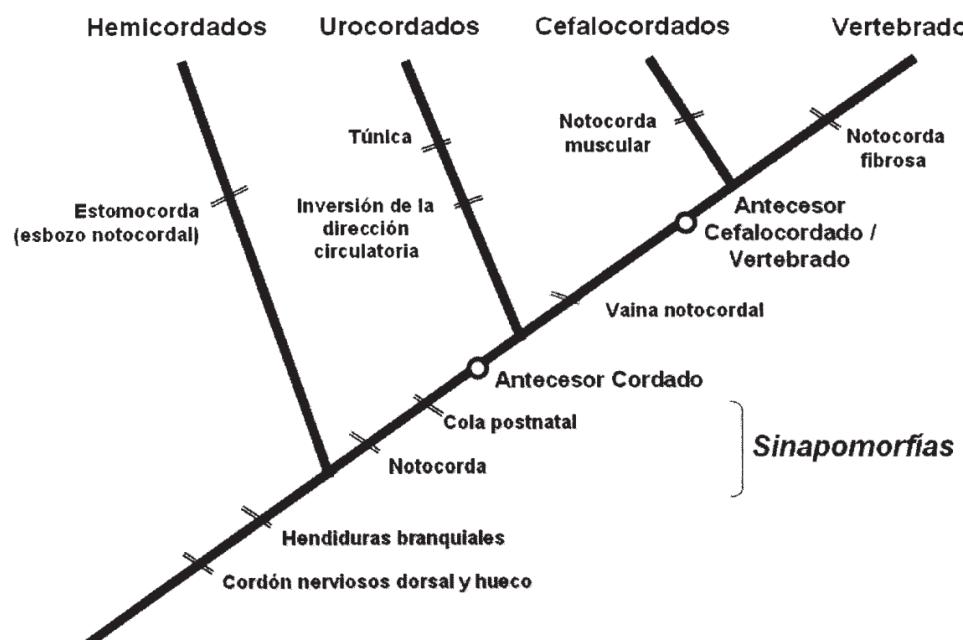
- **Symplesiomorfía** [Homología ancestral compartida - e.g. cuatro patas (tetrápodos)] se refiere a una plesiomorfía (carácter ancestral) compartida por dos o más taxa. They do not need to be associated

with monophyletic groups, could be paraphyletic or polyphyletic groups.

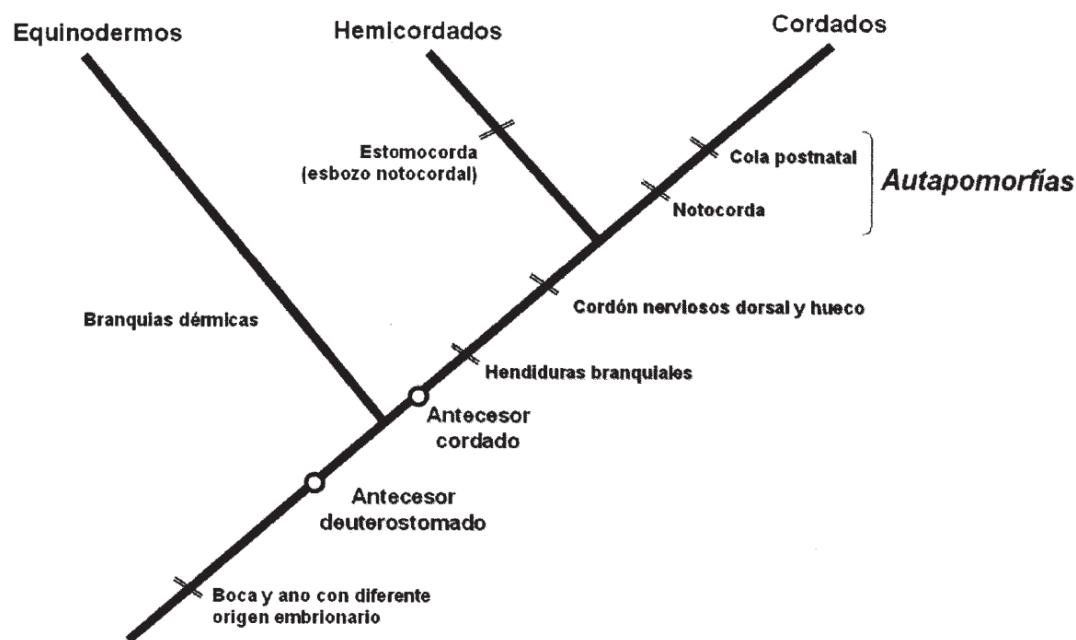


Apomorfía (derived character state) es un carácter novedoso evolutivamente y se dice que es derivado, ya que deriva de otro rasgo perteneciente a un taxón ancestral filogenéticamente próximo.

- **Sinapomorfía** [Homología derivada compartida - e.g. vision binocular (humano y mono)] una apomorfía (carácter exclusivo) compartida por un ancestro común y todos sus descendientes. Only synapomorphic character states are expected to be associated with *monophyletic groups* = clades.



- **Autapomorfía** [Homología derivada exclusiva - e.g. bipedismo (humano)] es un carácter novedoso y único de un taxón que no aparece en el antepasado, por lo que no lo comparte con ningún otro.



[*sýn-* gr. 'con', 'unión']

[*morph* gr. 'forma']

[*plesio-* gr. 'cercano']

+

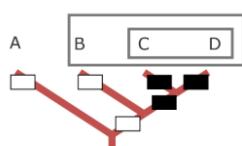
[*aut(o)-* gr. 'que actúa por sí mismo']

[*-ía* gr. 'cualidad']

[*apó-* gr. 'a partir de' (derivado, novedoso)]

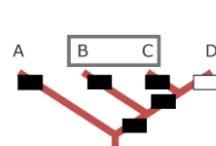
Grupos y caracteres que se apoyan: los grupos monofiléticos presentan sinapomorfía, los grupos parafiléticos presentan simplesiomorfía, y los grupos polifiléticos homoplasia.

Sinapomorfía



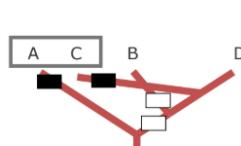
Grupo monofilético o clado

Simplesiomorfía



Grupo parafilético

Homoplasia



Grupo polifilético

Homoplasy

La homoplasia es una falsa homología, descubierta como tal a posteriori.

La homoplasia se refiere a la aparición de características similares en especies que no comparten un ancestro común reciente. Estas similitudes son el resultado de la **evolución convergente** o **paralela**, donde diferentes especies desarrollan rasgos similares de manera independiente como una adaptación a condiciones ambientales similares o a desafíos evolutivos parecidos.

- *Ejemplos* de homoplasia:

- Las alas de las mariposas y las de las aves son un ejemplo. Ambas sirven para volar, pero las estructuras y su origen son completamente diferentes. Una está soportada por exoesqueleto de quitina y la otra por un endoesqueleto óseo.

- La presencia de espinas en cactus y en ciertas plantas suculentas africanas es otro ejemplo de homoplasia.

La **analogía**, por otro lado, se refiere específicamente a la función o el rol de una característica, más que a su origen evolutivo. Dos estructuras son análogas cuando desempeñan funciones similares, pero no necesariamente comparten un origen evolutivo común. A menudo, las estructuras análogas son también homoplásicas, pero no siempre es el caso.

- Las alas de los murciélagos y las alas de las aves son análogas porque ambas se utilizan para volar, aunque tienen orígenes evolutivos diferentes.
- Las aletas de los pingüinos y las aletas de los delfines son estructuras análogas utilizadas para nadar.

En resumen, la homoplasia se centra en el origen evolutivo de las similitudes entre especies, mientras que la analogía se centra en la función de las características similares. Una estructura puede ser homoplásica y análoga al mismo tiempo si se desarrolló independientemente en diferentes linajes y cumple una función similar.

Hay 3 tipos de homoplasia en función de la causa que produce la homoplasia:

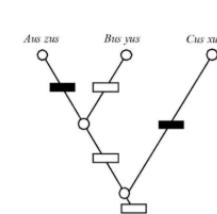
- La **convergencia** cuando la similitud de carácter se produce en grupos filogenéticamente lejanos. Se da cuando dos estructuras similares han evolucionado independientemente a partir de estructuras ancestrales distintas y por procesos de desarrollo diferentes.
- Se considera que el **paralelismo** cuando la similitud de carácter se produce en grupos filogenéticamente cercanos y anidados en un mismo clado; 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.

- Reversión** cuando un carácter se transforma en otro evolutivamente anterior, i.e. 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.



A



Fenotipo vs moléculas

Carácteres fenotípicos:

Tradicionalmente se han empleado para establecer las relaciones filogenéticas.

Pros:

- suelen ser caracteres **evolutivamente relevantes**
- caracteres complejos, **menos proclives a la homoplasia**
- son los únicos **caracteres disponibles** en algunos casos como en fósiles o especímenes raros

Contras:

- puede haber problemas de codificación de taxones supraespecíficos como terminales ("**quimeras**")
- se pueden dar casos de **subjetividad** en la codificación de caracteres
- hay un número **limitado** de caracteres fenotípicos
- podemos encontrar taxones altamente **autapomórficos** (exclusivos, debido a las muchas posibilidades evolutivas)

Carácteres moleculares:

Pros:

- son **estrictamente heredables**
- **no hay ambigüedades** en la codificación (determinar el estado de los caracteres es trivial)
- hay ciertas **regularidades en la evolución** de los caracteres moleculares
- son **robustos** frente a la distancia evolutiva
- son muy **abundantes**
- ofrecen información temporal: **reloj molecular**

Contras:

- son más proclives a la **homoplasia** al tener solo 4 nucleótidos y 20 aminoácidos
- la **evolución** de estos caracteres es **compleja**
- 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

Análisis separados

Se suelen utilizar multitud de **genes separados** y analizarlos aparte. El consenso de análisis separados es una estimación conservadora de la filogenia.

Conflictos entre caracteres:

- Algunos métodos filogenéticos sólo 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.

Análisis combinados

El conflicto entre caracteres se *resuelve* teniendo en cuenta:

- toda la **evidencia** disponible
 - 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.

Alineamiento de secuencias

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

Caracteres y estado del carácter

Caracteres son características del individuo que creemos que van a ser heredadas. Cuantos más caracteres estudiemos de cada individuo mejor.

Ej. un carácter que hayas adquirido a lo largo de tu vida no es heredable, en principio, por ej. el tamaño del cuádriceps no sería una buena referencia.

Estado de carácter es el valor específico que toma ese carácter en determinado taxón o especie.

Ej. carácter: ojo; estado: (numero de ojos) tener 8 ojos.

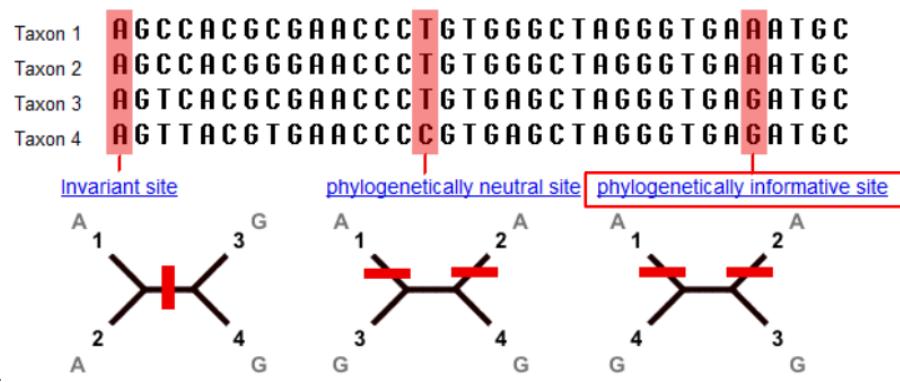
Un árbol filogenético se construye partiendo de la comparación de atributos (caracteres) que presentan variación entre los organismos objeto de estudio.

Un árbol se puede inferir a partir de diversos tipos de caracteres (morfológicos, moleculares, etológicos, ecológicos, biogeográficos, etc.), cuyas diferentes manifestaciones se denominan estados de carácter.

Existen dos tipos de caracteres en función de su *origen*: los *homólogos* y los *homoplásicos*.

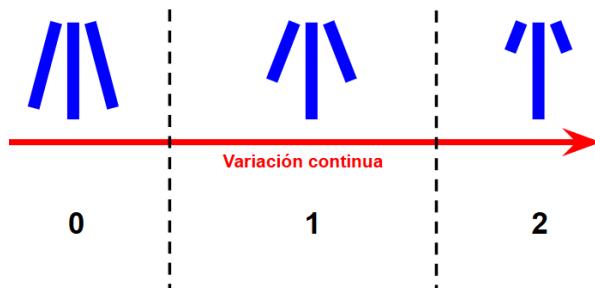
Tipos de caracteres

- Sitios **invariables**: que no cambian en los distintos taxones.
- Sitios **filogenéticamente neutrales**: que son autapomorfías (solo cambia en un taxón).
- Sitios **filogenéticamente informativos**: son comunes por pares (permiten dicotomía), son



Estados de un carácter

- **Binarios 0/1** (presentes o ausentes) [0 , 1]
- **Multiestado o binarios V/S** (transversiones o transiciones)
- **Discretos o continuos:**
 - Discretos: Ej. Número de dedos
 - Continuos: Ej. Diferencia de longitud (%) entre el dedo medio y los otros; diferencia entre la longitud del pico. 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.



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.

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.

Ponderación a posteriori

El método más utilizado y aplicado, también llamado **ponderación implícita** 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 homoplásia se hace más evidente. Los árboles que maximizan la función cóncava de homoplásia resuelven el conflicto de caracteres a favor de los caracteres que tienen más homología (menos homoplásia) 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.

$$F = \sum f_i, f_i = k/k+(s-m)$$

m = minimum number of steps

s = number of steps observed

k = concavity constant, 1-100

$(s-m)$ mide el número extra de cambios (homoplásia)

Un valor de k próximo a 0 prácticamente elimina del análisis los caracteres más homoplásicos. Por contra, valores elevados ponderan negativamente de forma muy suave los caracteres homoplásicos.

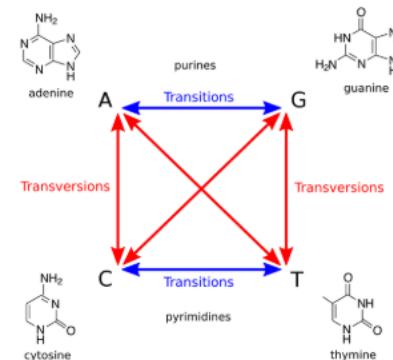
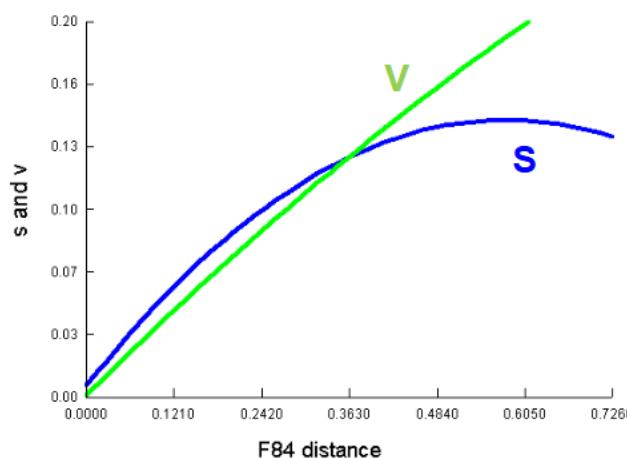
morfología: $k=5-16$

secuencias nucleotídicas: $k>15$

Secuencias de ADN

Generalmente, se toma la tasa de sustitución como medida de la fiabilidad de la información filogenética del marcador. Se entiende entonces homoplásia como saturación

$\$Homoplásia = Saturación\$$



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.

Hacer una transición es más difícil que una transversión, por ello si una transición ocurre es más importante, i.e. las transiciones (S) son más costosas que las transversiones (V)

- **Peso de los caracteres:** Debido a cómo se forman los aminoácidos, los cambios en la tercera posición son menos importantes (tienen menos peso) que los de la primera.

Polaridad de los caracteres

Para orientar un árbol en el tiempo o enraizarlo es necesario determinar qué estados son plesiomórficos y cuales apomórficos (es decir, polarizar los estados).

Utilizando un **criterio ontogenético**, se ve cómo se forma el carácter durante el desarrollo (desarrollo del individuo, referido en especial al período embrionario) 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.

De todos los caracteres presentes en un grupo monofilético, aquel que se encuentre en su grupo hermano, corresponderá al carácter plesiomórfico, mientras que el que se encuentre exclusivamente en el grupo interno será el apomórfico.

Homología de los caracteres moleculares

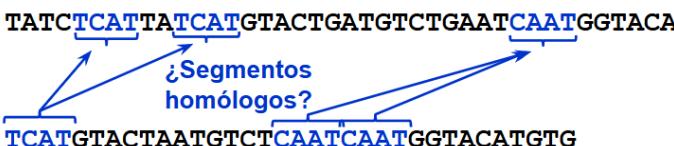
Cuando analizamos secuencias, asumimos que son de moléculas heredadas de ancestros a descendientes (**ortólogos**).

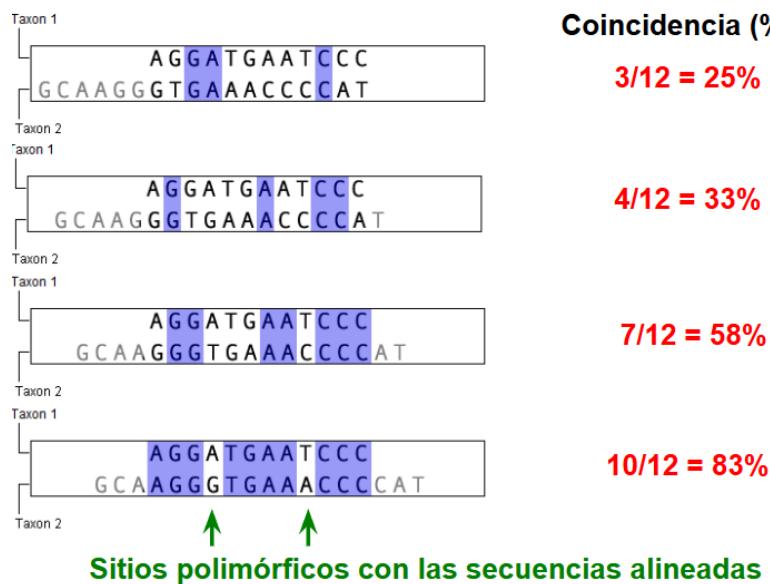
Cada secuencia está formada por muchos caracteres (cada posición en la secuencia). Por ello, **un primer paso es determinar el estado** de cada uno de esos caracteres en cada taxón de la matriz.

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.

El concepto de Homología aplicada 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.



- **Gaps** son marcadores de posición que introducimos en los alineamientos para mantener la homología posicional (para secuencias de distinta longitud). Representan eventos de inserción o pérdida denominados *indels* (del inglés insertion/deletion).
 - Pros: son, en principio, menos propensos a la homoplasia que las sustituciones de bases, muy utilizadas en análisis de parsimonia.
 - Contras: son difícilmente gestionables por la mayoría de los modelos de evolución molecular.

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

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.
- 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.

Pero como los gaps no penalizan pueden darse alineamientos con misma puntuación, pero más posiciones de diferencia.

Para solucionar esto se pueden aplicar **penalizaciones para los GAPS**:

- penalizaciones por la apertura de los huecos

- penalizaciones por la extensión de los huecos abiertos (típicamente menores que las impuestas por

<table border="1" style="border-collapse: collapse; width: 100%; text-align: center;"> <tr><td>A</td><td>T</td><td>C</td><td>G</td></tr> <tr><td>A</td><td>T</td><td>T</td><td>G</td></tr> <tr><td colspan="4">$1+1+0+1 = 3$</td></tr> </table>	A	T	C	G	A	T	T	G	$1+1+0+1 = 3$				<table border="1" style="border-collapse: collapse; width: 100%; text-align: center;"> <tr><td>A</td><td>T</td><td>-</td><td>C</td><td>G</td></tr> <tr><td>A</td><td>T</td><td>T</td><td>-</td><td>G</td></tr> <tr><td colspan="5">$1+1-2-2+1 = -1$</td></tr> </table>	A	T	-	C	G	A	T	T	-	G	$1+1-2-2+1 = -1$				
A	T	C	G																									
A	T	T	G																									
$1+1+0+1 = 3$																												
A	T	-	C	G																								
A	T	T	-	G																								
$1+1-2-2+1 = -1$																												

apertura). **3 > -1 el de la izquierda es mejor**

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.

Una vez alineadas las secuencias...

Tratamiento de gaps:

- Como datos perdidos (missing data). Es la opción más usada por la mayoría de programas de filogenia.
 - Inconveniente: Descarta información utilizada para definir la homología de las demás posiciones.
 - Los gaps pueden proporcionar importante información filogenética.
- Gaps como quinto estado (A, T, G, C, Gap). Inapropiado, porque:
 - Los gaps son el resultado de una forma diferente de cambio (problemas con los modelos).
 - Los gaps de más de una posición incluyen caracteres tratados como independientes unos de otros, aunque puedan resultar todos de un único evento de indel (se puede ponderar).
- Codificación de gaps como presencia/ausencia.
 - No aumenta la ponderación de los varios gaps no homólogos que se solapan.
 - Los indels son utilizables en parsimonia, pero también en métodos de inferencia Bayesiana.

Existen reglas para recodificar los gaps, basadas en su solapamiento y la compartición de sus extremos 5'y/o 3'

- Simple: 2xread, GAPCODER, GAPRECODER, FASTGAP
- Complejo: SEQSTATE

Modelos de evolución molecular

Supuestos de los métodos de reconstrucción filogenética

Todos los *métodos de inferencia* asumen que (aunque éstos no se hagan explícitos):

- Todos los **sítios 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 las asunciones, pero en realidad no son ciertas...

NOTE:

- las posiciones NO cambian independientes las unas de las otras,
- Las tasas de evolución NO son constantes,
- 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.

Se intentan hacer correcciones a esto, aplicando los **modelos de sustitución**. A nivel probabilístico se hacen predicciones.

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, i.e. nucleótidos o aminoácidos) cambian a lo largo del tiempo.

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

$$\text{Modelo} = \boxed{\pi = [a,c,g,t]} + \boxed{\begin{array}{|c c c c|} \hline a & b & c & d \\ \hline b & a & e & f \\ \hline c & e & a & g \\ \hline d & c & f & a \\ \hline \end{array}} + \boxed{\text{Diagrama de transiciones}}$$

- **Composición:** frecuencia de las diferentes bases o aminoácidos. La frecuencia de los nucleótidos se representa por:

$$\$ \pi = [0.25 \ 0.25 \ . \ .] \$$$

Se estima a partir de los datos.

Si trabajamos con proteínas, sería la frecuencia de aminoácidos.

- **Proceso de sustitución:** tasa de cambio de uno a otro estado de carácter.

$$P = \begin{matrix} \begin{array}{cccc} \textcolor{blue}{A} & \textcolor{blue}{C} & \textcolor{blue}{G} & \textcolor{blue}{T} \\ \hline \end{array} \end{matrix} \left(\begin{array}{cccc} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{array} \right)$$

El proceso de sustitución se representa mediante una matriz.

- Para secuencias de nucleótidos, hay 16 cambios posibles (una matriz de 4 x 4).

- Para los *nucleótidos*, se puede estimar a partir de los datos. Por ejemplo:

$$P = \begin{pmatrix} a & c & g & t \\ 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{pmatrix}$$

- La probabilidad de que una "a" cambie por una "c" es 0.01, la probabilidad de que una "c" se mantenga como está es 0.983, etc.
- Las filas de la matriz suman 1 (se cubren todas las posibilidades para cada nucleótido)
- Las columnas no suman nada en particular
- Para los *aminoácidos*, la matriz Q es fija y no tiene ningún parámetro libre. Hay diferentes modelos: Dayhoff y JTT (DNA nuclear), mtREV24, mtMAM, mtART (mtDNA) y BLOSUM 62 y WAG (secuencias de aminoácidos distamente relacionadas). Para los diferentes modelos, las tasas de sustitución de aminoácidos se estiman a partir de datos empíricos.

- Otros parámetros (heterogeneidad de tasas):** proporción de sitios invariables o agregación de los cambios a lo largo de la secuencia.

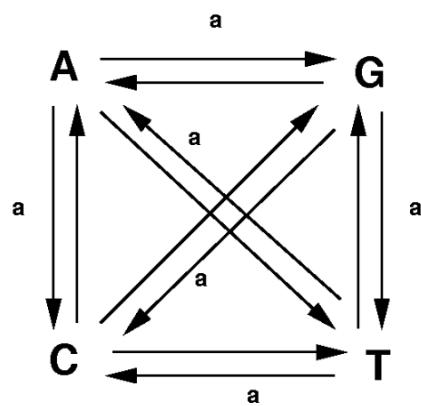
Modelos frecuentes

Algunos de los modelos más frecuentes son:

1. Jukes and Cantor (JC69):

El modelo más sencillo, asume que:

- 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 (todos los cambios son igualmente probables).



La complejidad de los modelos va en aumento, ya que las combinaciones de parámetros son muchas...

2. 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.

3. Hasegawa-Kishino-Yano (HKY):

Como K2P, pero la composición de bases varía libremente.

4. 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.

Heterogeneidad de tasas de sustitución

Los *modelos anteriores* asumen que:

- el cambio es igualmente probable en todas las posiciones de la secuencia y
- la tasa de cambio es constante a lo largo de la filogenia.

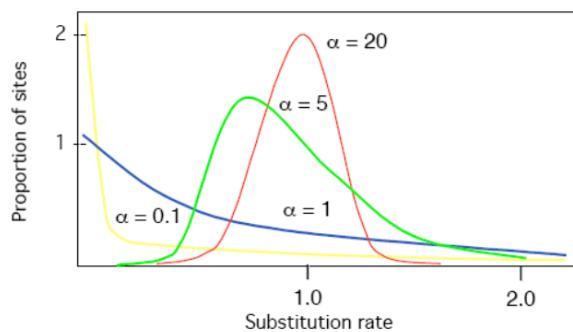
Pero la intensidad de la selección es rara vez uniforme a lo largo de las posiciones, de modo que lo deseable es **modelar la variación de las tasas de sustitución sitio por sitio**.

¿Cómo se modela dicha **heterogeneidad**?

- 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**.

- También esperamos observar **posiciones VARIABLES**, 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.

Las tasas de sustitución se calculan para el conjunto de datos. Porque lo que hacemos es compararlos entre ellos.

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.

$$L = \Pr(D|H)$$

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 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).

Selección de esquemas de partición

Trata de elegir la mejor forma de **particionar los datos**.

- 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](#).

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**: modelo de sustitución y esquema de partición

6. **Inferencia filogenética:** MP (máxima parsimonia), minimum evolution, neighbour-joining, ML (máxima verosimilitud), inferencia Bayesiana

7. Construcción del **árbol filogenético**:

1. **Soporte estadístico:** Non-parametric bootstrap, posterior probabilities
2. **Testar hipótesis filogenética:** Parametric bootstrap, Kishino-Hasegawa, Shimodaira-Hasegawa, approximately unbiased
3. **Estimación del tiempo de divergencia:** Reloj molecular (convencional), Penalized likelihood, Bayesian rate autocorrelation dating, Bayesian uncorrelated relaxed clock

Busqueda de árboles

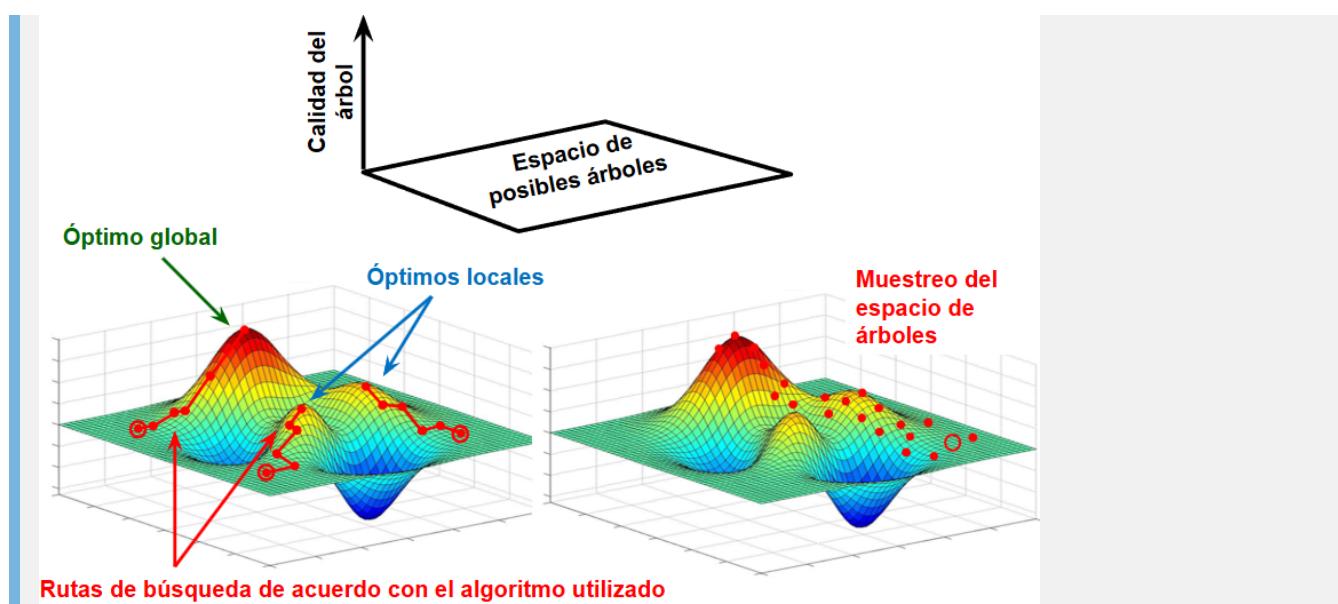
Sólo hay una manera de construir el primer árbol sin raíz, uno con tres puntas (nodos terminales) y tres ramas. Cada vez que añadimos un taxón, se crean dos ramas.

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*" 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).



Los árboles en la zona más alta son los mejores (con menos pasos y que maximizan la verosimilitud de los datos).

Cuando se empieza la búsqueda no se sabe donde estás. Pero según analizas va aumentando el likelihood de que consigas el árbol óptimo.

Mientras más corras los datos, más posibilidad hay de que encuentres el óptimo

Algoritmos de reordenación de ramas (branch swapping):

A. Nearest Neighbour Interchange (NNI):

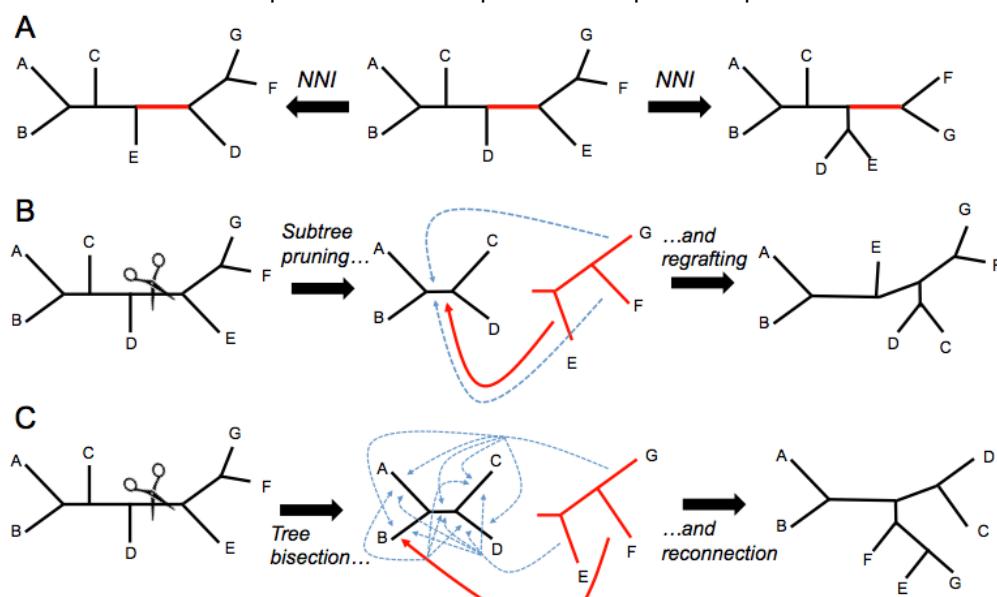
intercambia dos vecinos por cada rama interna.

B. 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.

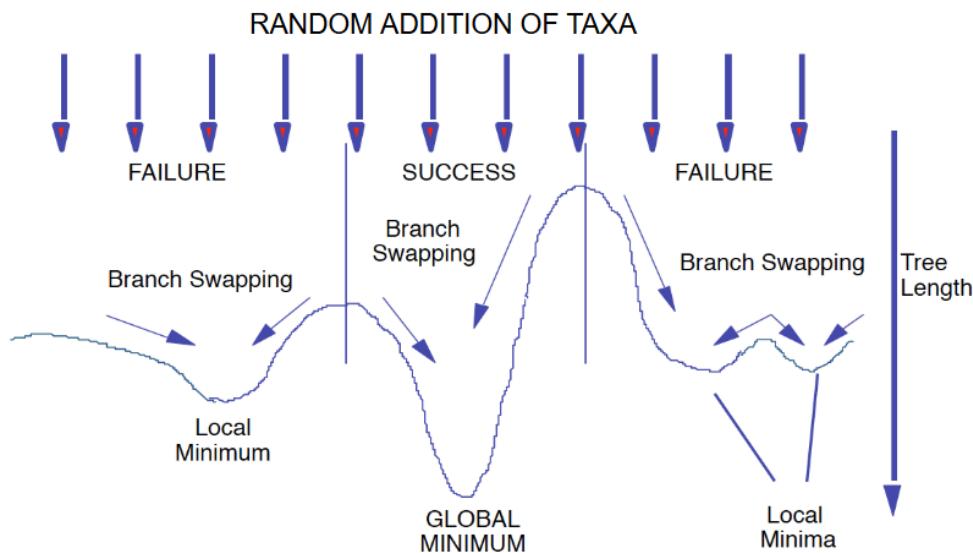
C. Tree Bisection and Reconnection (TBR):

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.

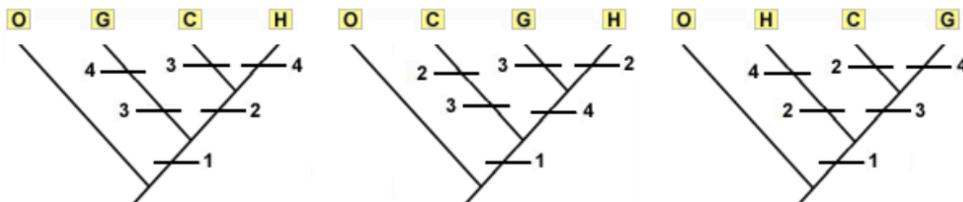
El espacio de árboles puede estar poblado por mínimos locales e islas de árboles óptimos.



Árboles de consenso

A menudo existen varios árboles candidatos a ser el cladograma más parsimonioso.

En el siguiente ejemplo, hay tres cladogramas diferentes que son igualmente 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 de resumir los puntos de acuerdo entre cladogramas rivales para llegar a formar un "árbol de consenso".

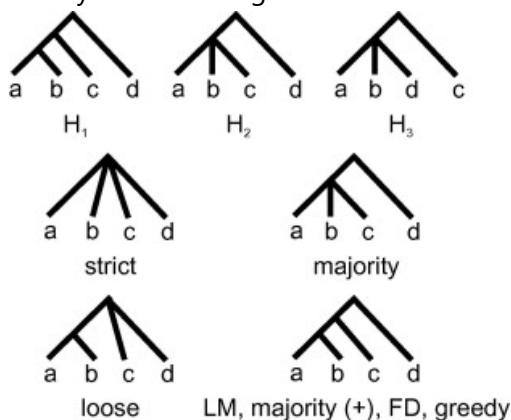
Un **árbol de consenso** es un árbol que combina los *agrupamientos preferidos* a partir de los cladogramas rivales de un determinado grupo de taxones, de tal forma que los agrupamientos discutibles (contenciosos, ambiguos) se condensan en puntos con múltiples ramas (politomías).

Existen diferentes **formas para construir árboles de consenso**, pero los tres métodos más comunes son:

1. **Árbol de consenso estricto:** conserva sólo los agrupamientos que comparten todos los cladogramas rivales.

2. **Árbol de consenso semi-estricto:** conserva todos los agrupamientos que no son contradictorios en los cladogramas rivales.

3. **Árbol de consenso de regla de la mayoría:** conserva todos los agrupamientos que son apoyados por la mayoría de cladogramas rivales.



Medidas de soporte: confianza en el árbol

La mayor parte de las medidas científicas van acompañadas de una estima de su precisión.

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:

1. 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.

2. Error sistemático:

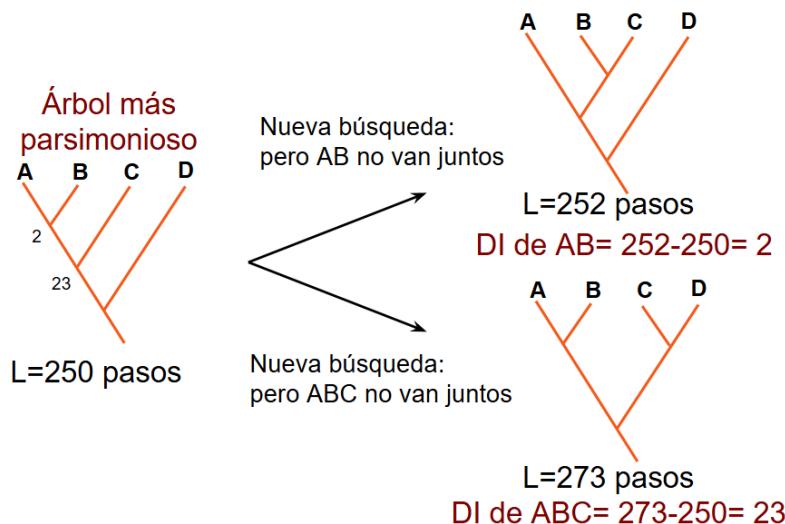
Asociado a la metodología y las asunciones de los análisis.

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).

Esta información se puede dar en tablas adicionales para mostrar porque tu confianza.

1. Soporte de Bremer o Decay Index:

Apoyo *cualitativo* donde 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.

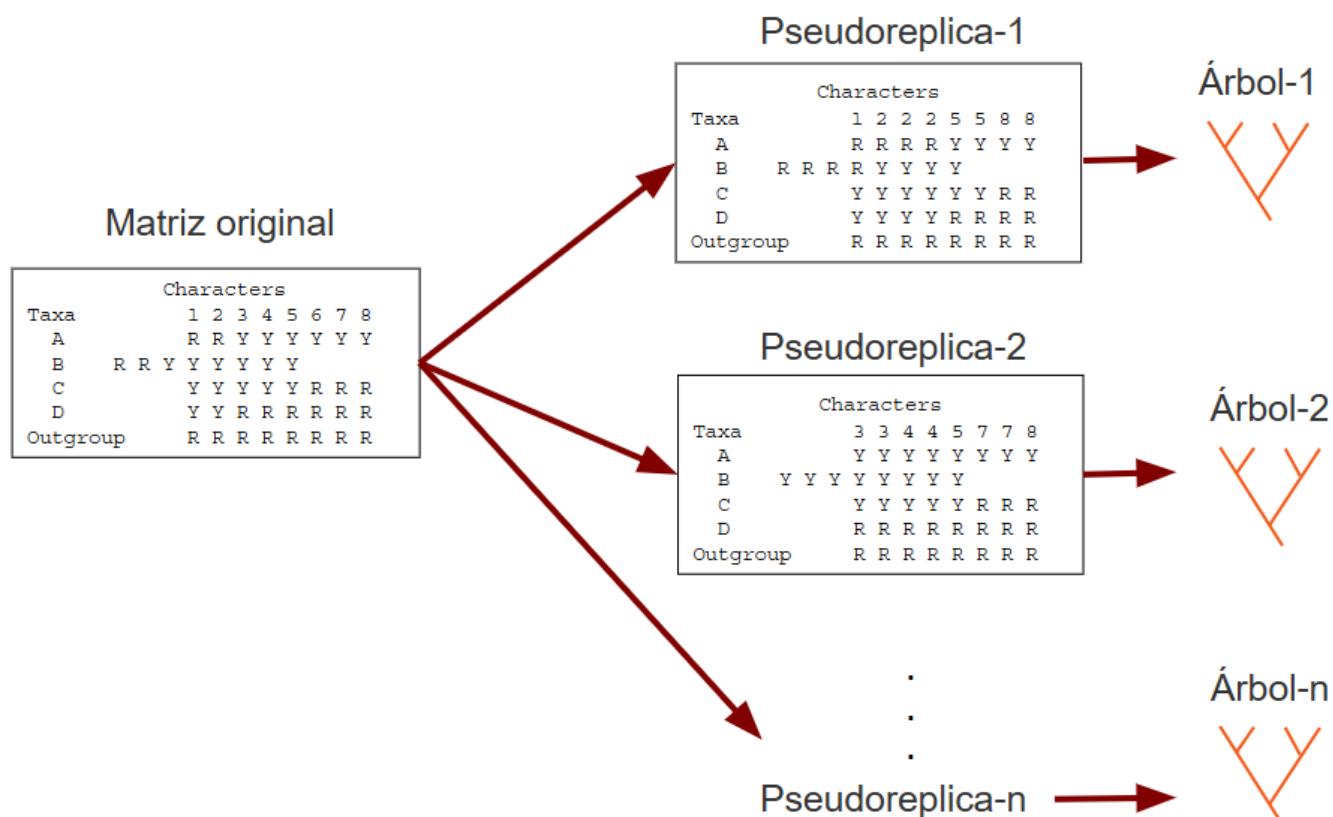


2. Remuestreo por bootstrapping:

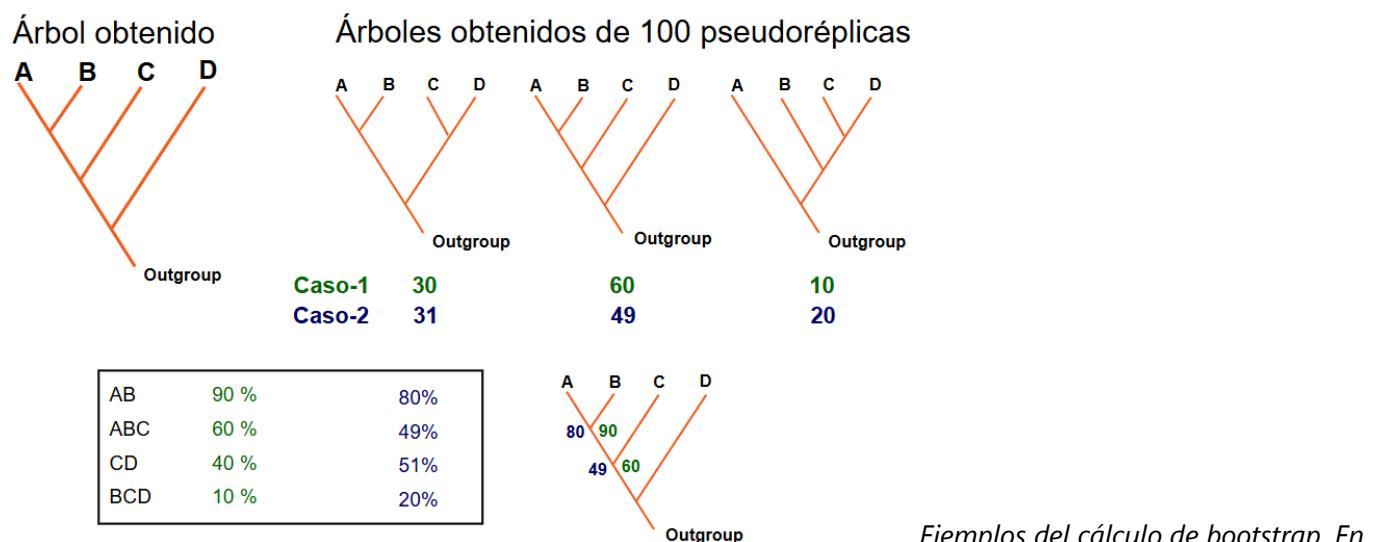
Se remuestrean los caracteres al **azar**, con **reemplazamiento, múltiples veces** (entre 500 y 2000, normalmente 1000)

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 pseudoreplicas se construyen a partir de la matriz original con reemplazamiento para construir una nueva matriz del mismo tamaño que la original.

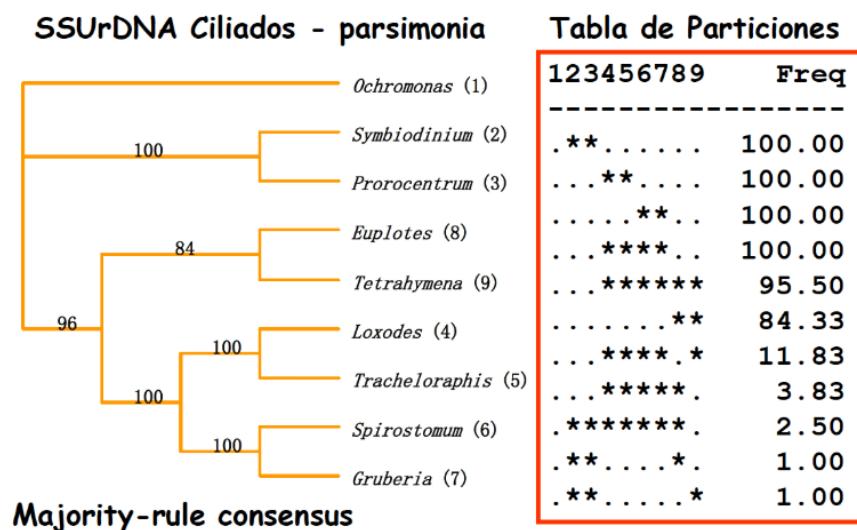


Ejemplos del cálculo de bootstrap. En

el primer caso (mostrado en verde), tras 100 pseudorélicas, 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.

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 probabilities** (BPs) son conservadores.

Los BPs 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 BP 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 aumentar, según se considere.

3. 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 pseudoreplicas 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.

Máxima parsimonia (MP) - 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.

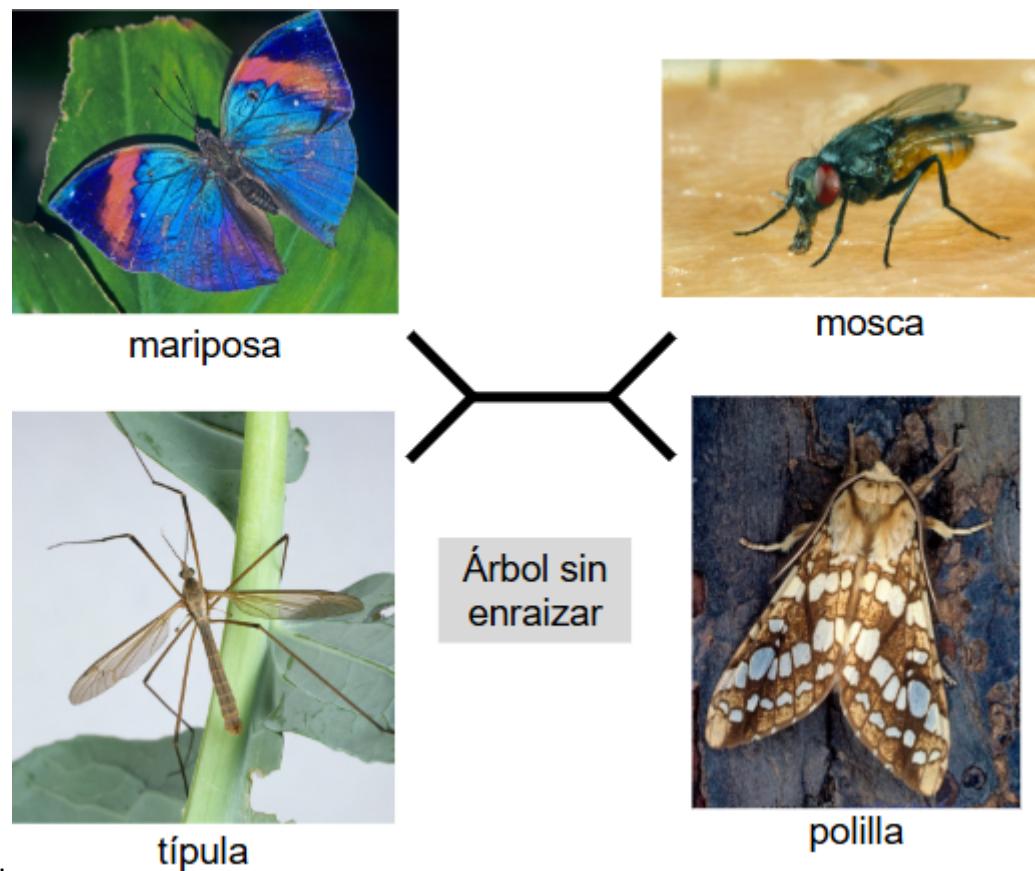
Descansa sobre el *axioma fundamental* de que en la naturaleza, como resultado de la evolución, existe un orden que se manifiesta en las **similitudes de los caracteres**.

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)**

Paso 1: elegir el arbol más parsimonioso

Desconocemos el aspecto del antecesor común más reciente de las siguientes 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...

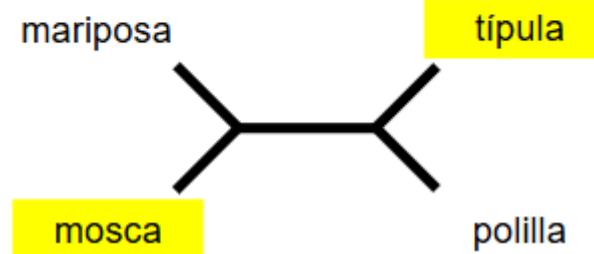
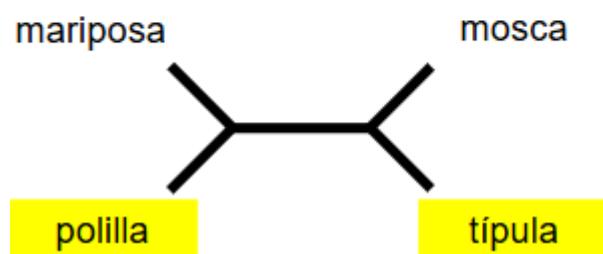
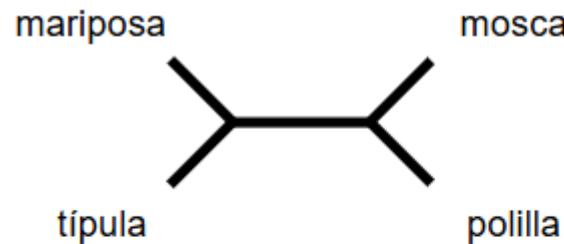
Este ejemplo implica mayor similitud entre...

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 puede obtener otra red cambiando la posición de las especies...

Existen tres posibles formas de unir 4 especies en una red de ramificación dicotómica.

¿Cuál es la red que explica de forma más sencilla la distribución de los diferentes estados de caracteres que distinguen las cuatro especies?

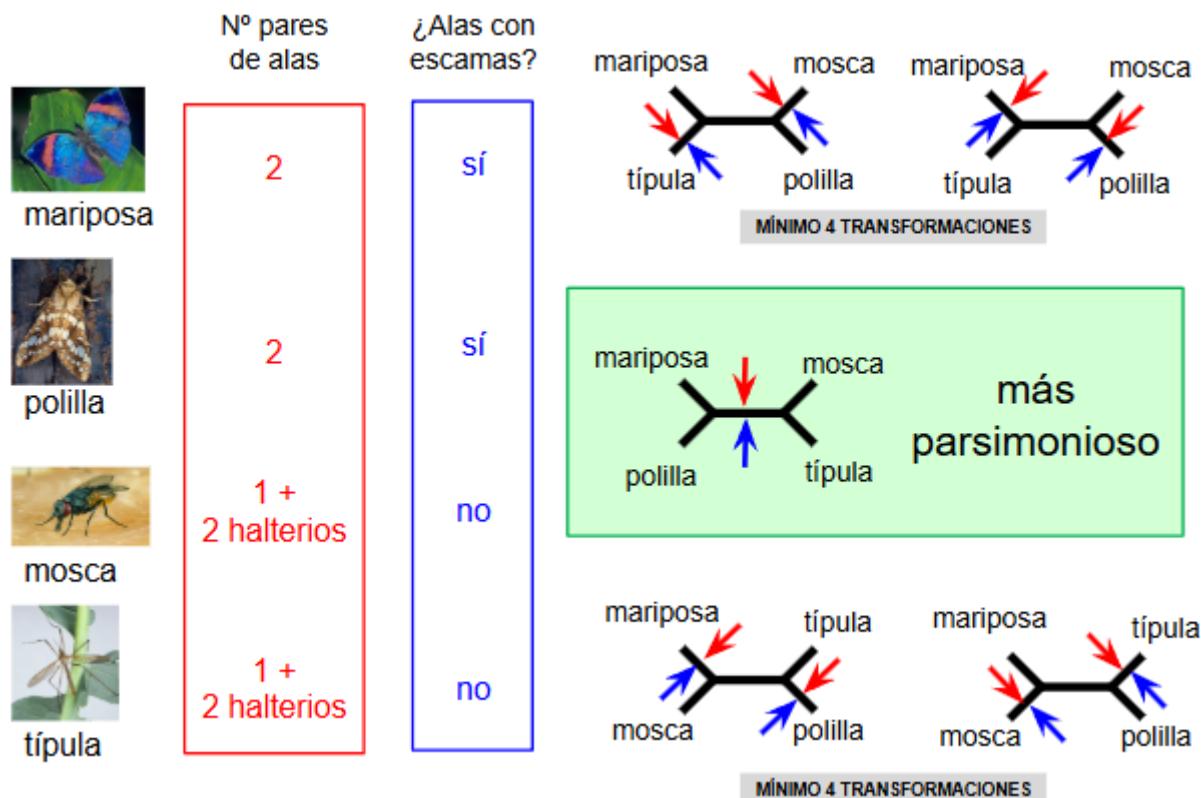
*La selección de la explicación más sencilla se denomina **principio de parsimonia**.*

Pasos a seguir:

1. Revisamos los caracteres de las 4 especies para ver si pueden ayudarnos a elegir entre los tres árboles



2. Construimos la matriz de caracteres... y a continuación miramos cómo se sitúan los estados de los caracteres en las tres redes posibles



En cada árbol de arriba y abajo del todo vemos que han habido 4 transformaciones, para que los caracteres dados sean posibles.

El árbol del centro, muestra que solo una vez se ha ganado o perdido los caracteres y esto hace que este sea el árbol más parsimonioso.

3. Aplicando el principio de parsimonia rechazamos el primer y tercer árbol y nos quedamos con él segundo como el **más parsimonioso** ya que "es el que necesita menos transformaciones evolutivas para explicar la distribución de los estados de los caracteres analizados".

Principio de parsimonia

Ante distintas hipótesis sobre las relaciones filogenéticas de los organismos dados, elegimos la más parsimoniosa, esto es, la que tenga:

- **longitud menor** (menor número de cambios = menor numero de pasos evolutivos)
- **mayor número de homologías y**
- **menor número de homoplasias**

"la explicación más sencilla es la verdadera"

Si hay varios árboles o hipótesis igualmente parsimoniosos, no podremos elegir entre ellos.

Longitud de un árbol (L)

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.

Paso 2: Enraizar el arbol

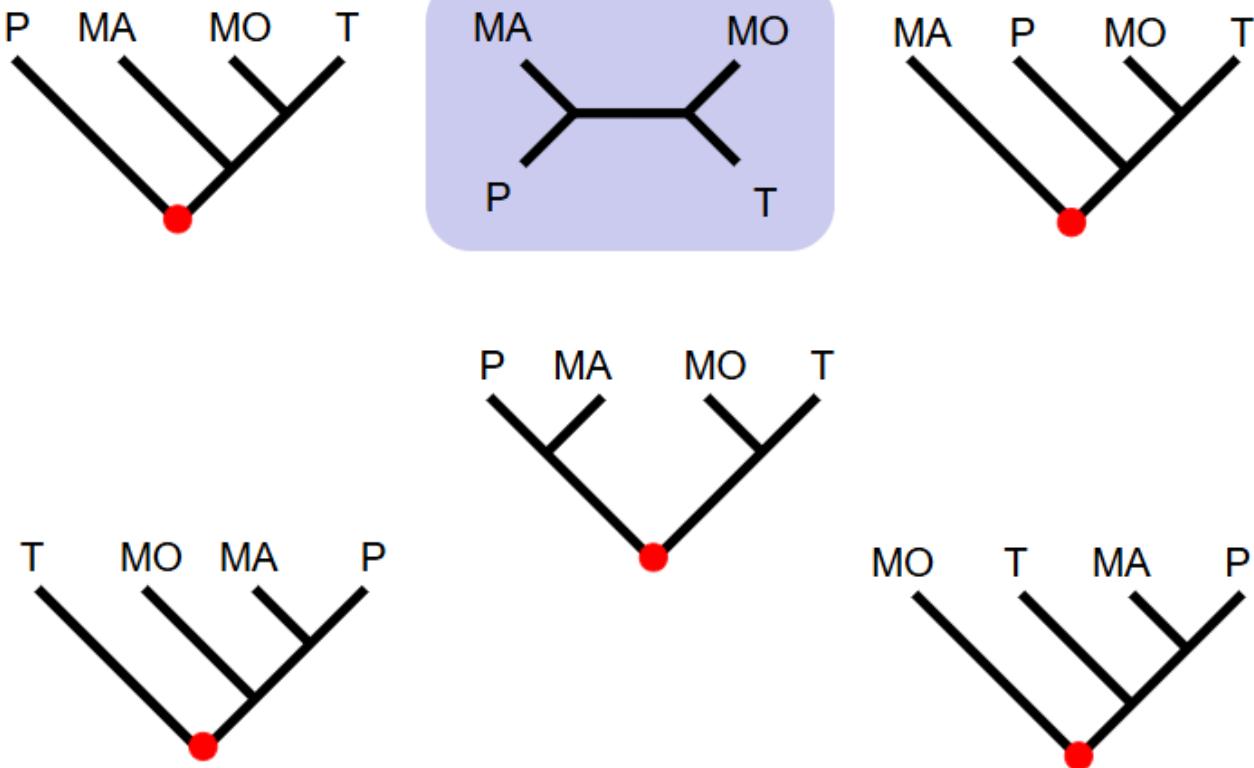
El siguiente paso es enraizar la red, la raíz nos permite determinar el lugar donde se encuentra el ancestro común.

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).

Siguiendo el ejemplo anterior.

Como partíamos de 3 redes diferentes, cada una de las cuales podría dar lugar a 5 cladogramas, en total tendríamos 15 cladogramas diferentes para representar las relaciones evolutivas de 4 especies... Sin embargo, como previamente descartamos dos de las redes por necesitar de un mayor número de transformaciones evolutivas (y ser menos parsimoniosas), nos quedamos con los cinco cladogramas provenientes de la red más parsimoniosa.



5 posibilidades de enraizar una red con 4 especies

...hay varias posibilidades... supongamos por ejemplo que el antecesor se encontraba en la rama de la polilla...

...obtenemos un primer cladograma.

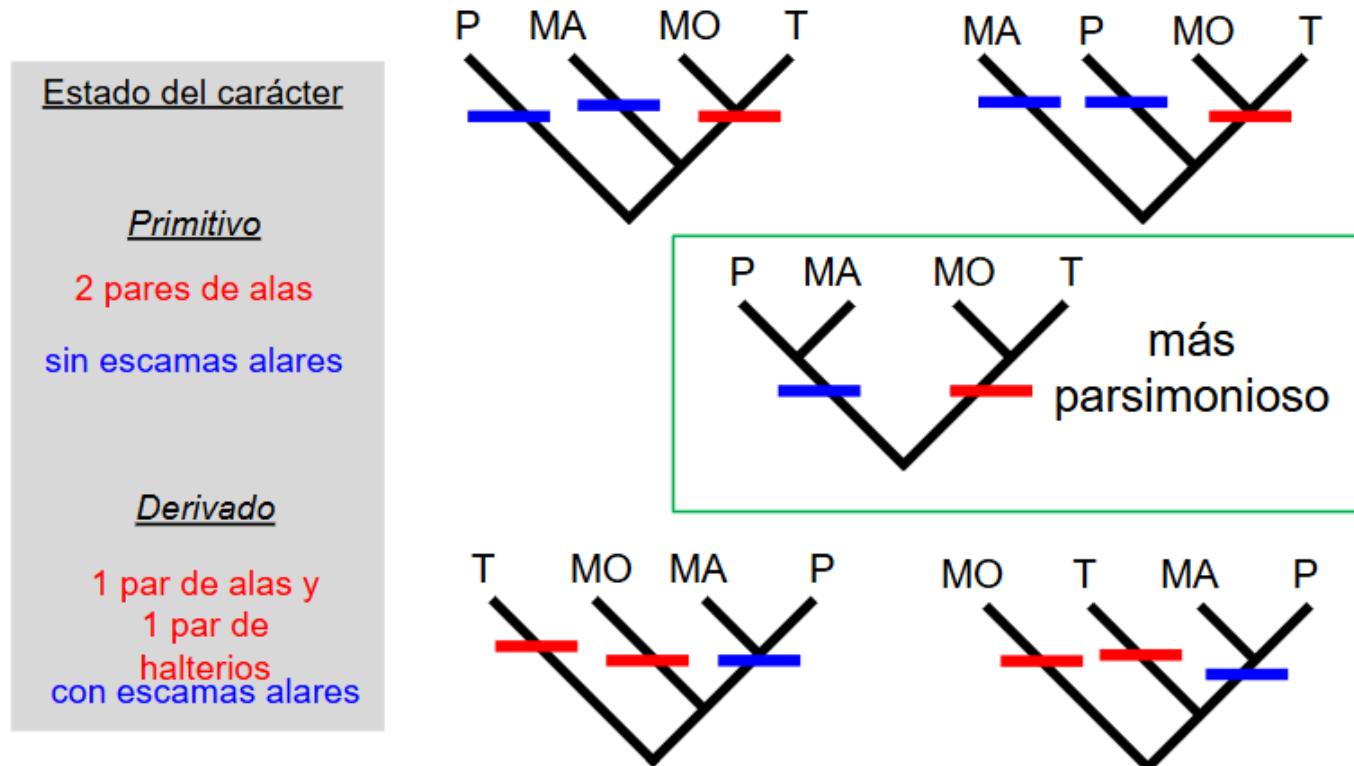
Como en la red hay otras 4 ramas, podemos construir 4 nuevos cladogramas...

Ahora se trata de ver cuál de ellos es el más probable aplicando el principio de la parsimonia a la distribución de los estados de los caracteres.

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.

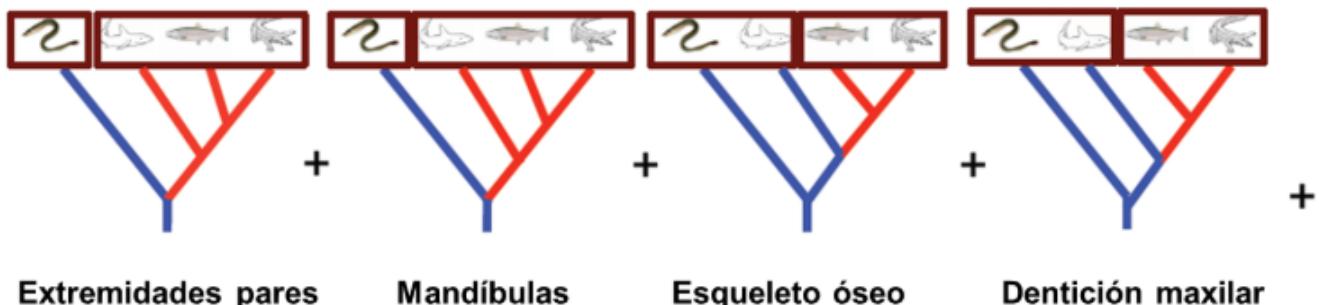


El árbol más parsimonioso es aquel que necesita menos transformaciones para explicar la distribución de los estados derivados de los caracteres, y el que probablemente mejor registre las relaciones evolutivas de las cuatro especies de insectos.

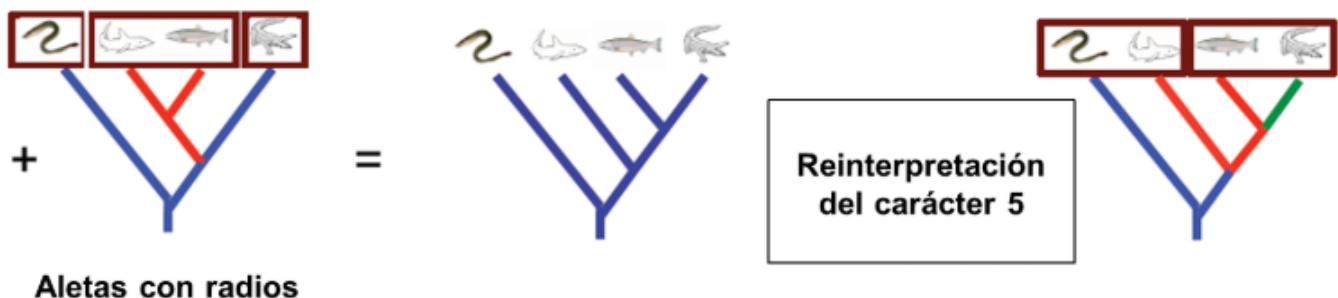
La máxima parsimonia asume los siguientes principios:

- **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 **parallelismos** 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*.



Homología primaria



Homología secundaria

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** (homoplásia).

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.

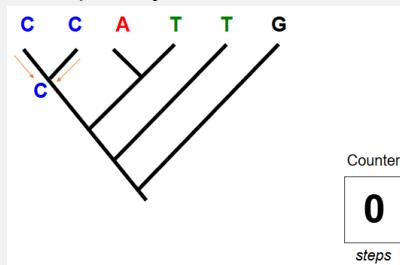
Optimización: Evaluación de cladogramas - Algoritmo de Fitch

Para decidir el número de pasos, ir de arriba abajo y después volver a analizar de abajo a arriba, para adivinar donde han ocurrido los cambios y resolverlo con el menor número de pasos posibles.

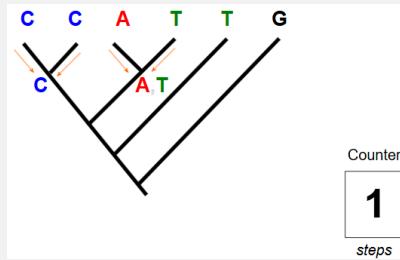
Esto se hace de la siguiente manera:

1. Calcular el número de pasos asumiendo que siempre hay homología. Yendo de arriba a abajo del árbol:

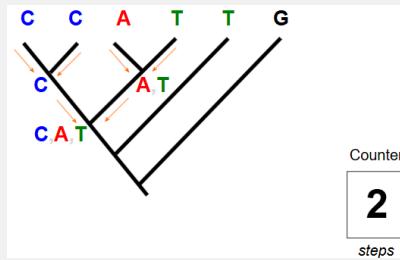
1. Para que haya dos C, el ancestro común debe tener una C



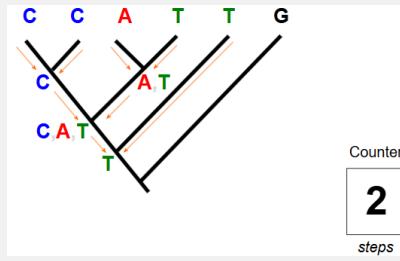
2. A, T tendría A y T el ancestro común



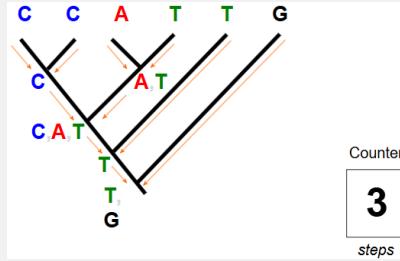
3. El sig nodo tiene C,A,T



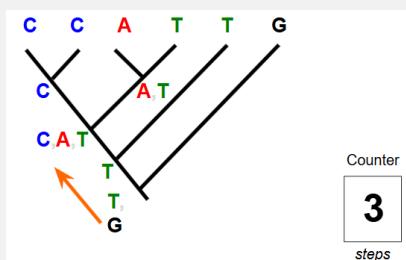
4. El siguiente, tiene una T, porque es la común entre ambas



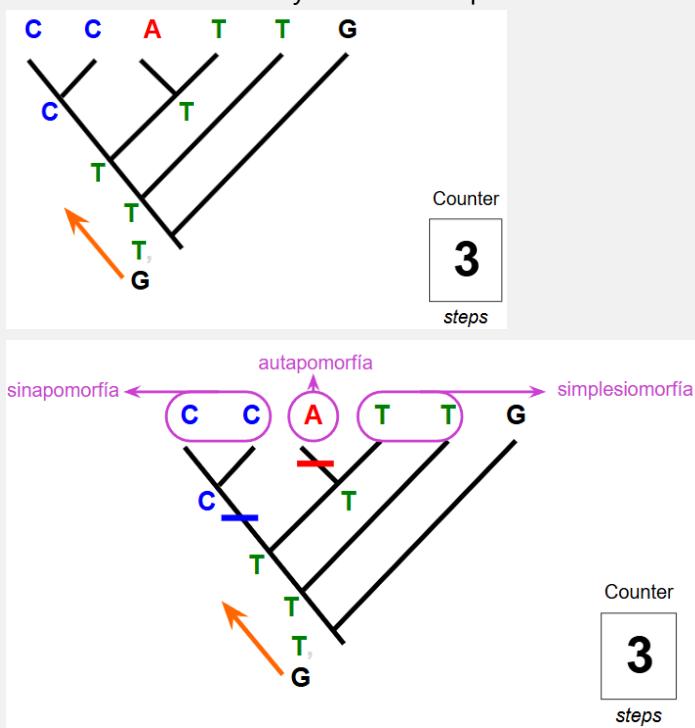
5. La última es T, G porque al ser distintas, se mantienen ambas



2. Seguir el árbol de abajo a arriba: observar lo más parsimonioso siguiendo el árbol hacia arriba.

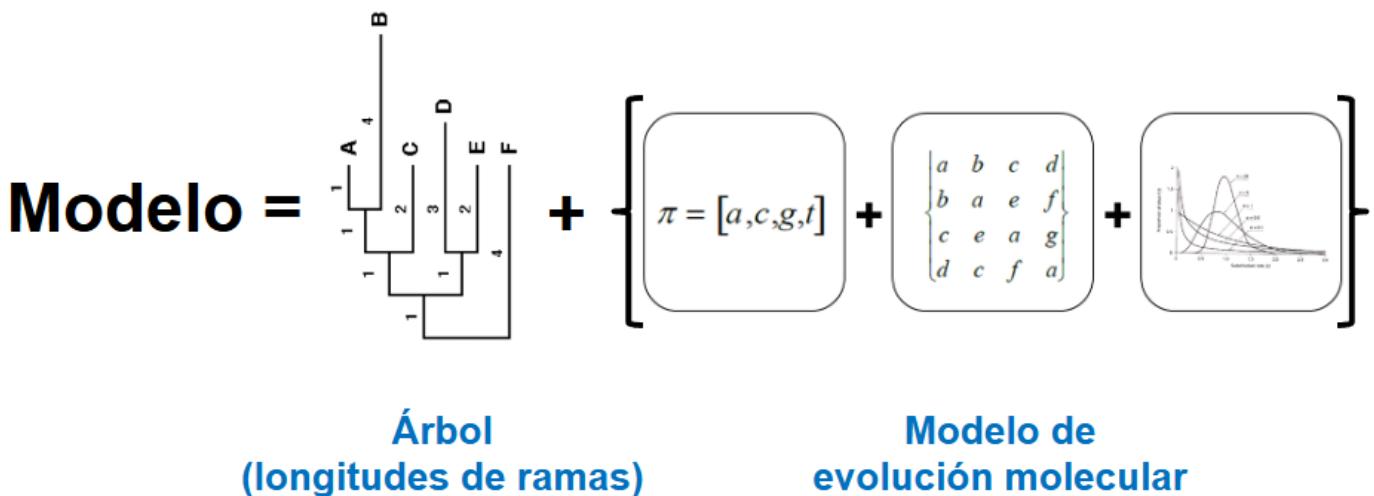


- Por número de cambios, lo más parsimonioso es que la T se mantiene, y la novedad evolutiva es la C y la única autapomorfía es la A.



Máxima verosimilitud (ML)

Los métodos probabilísticos se apoyan en la verosimilitud de obtener los datos (un alineamiento múltiple de secuencias) si los linajes hubieran evolucionado de acuerdo con un determinado árbol filogenético (con su topología y longitudes de las ramas) y bajo un determinado modelo de evolución molecular.



La máxima verosimilitud (ML) intenta responder a la siguiente pregunta: ¿cuál es la **probabilidad de observar los datos** (un alineamiento de secuencias), **dada una hipótesis** (un árbol y un modelo concreto de evolución molecular)?

$$L = \Pr(D|H)$$

L: Verosimilitud del modelo (un árbol filogenético), es igual a la probabilidad (*Pr*) de los datos (*D*) (un alineamiento de secuencias) dada una hipótesis (*H*) (un árbol y un modelo de evolución molecular).

El árbol que hace que los datos sean el resultado más probable es una estimación de máxima verosimilitud de la filogenia. Se hacen dos estimaciones:

1. ¿Cuál de las posibles topologías hace los datos más verosímiles? (NNI, SPR, TBR)
2. Para una topología: ¿qué longitudes de ramas hacen los datos mas verosímiles?

Esto es una consideración filosófica, ya que la verosimilitud calculada *no es la probabilidad de que el árbol sea el correcto*, sino la **probabilidad de que el árbol estimado generase los datos** (si cambian los datos, cambia el árbol). En otras palabras, la ecuación de verosimilitud no es la probabilidad de que la hipótesis sea correcta en términos absolutos, sino para nuestros datos.

La máxima verosimilitud cuenta con una serie de supuestos:

- Los sitios evolucionan independientemente
- Los cambios siguen un modelo de Markov: la probabilidad de que tenga lugar un cambio en un sitio no depende de la historia previa de ese sitio
- Los cambios son reversibles en el tiempo

El procedimiento interno calcula la verosimilitud de un alineamiento de dos secuencias, dada una matriz de sustituciones, cierta composición de bases y una longitud concreta para la rama que separa esas secuencias (CED: **certain evolutionary distance**).

Para ello, se calculan todos los eventos posibles.

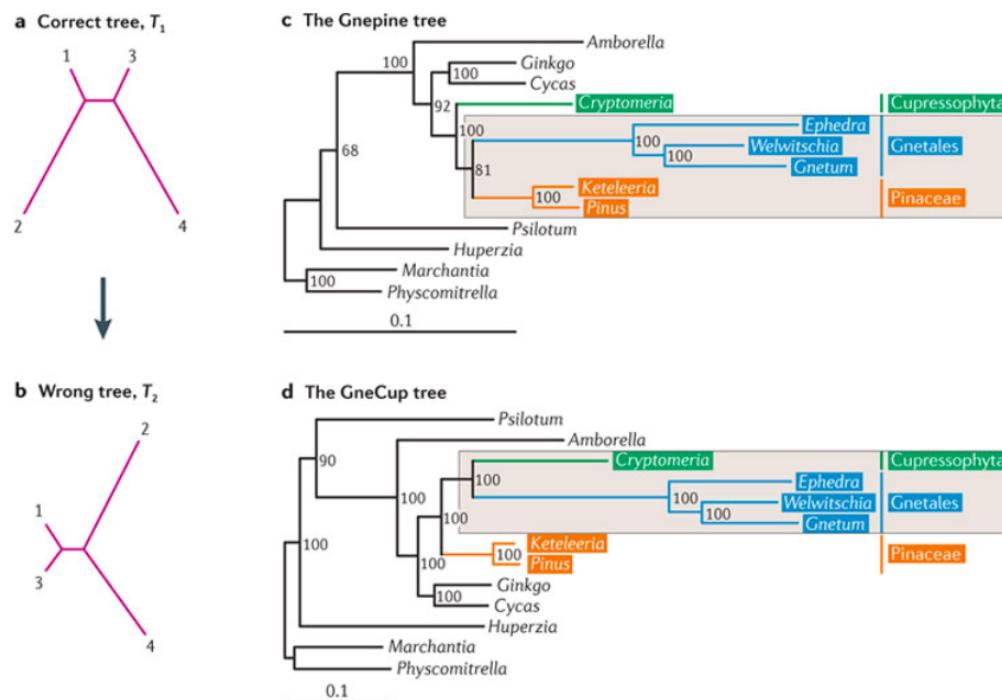
$$\begin{array}{l}
 \text{ccat} \\
 \downarrow \\
 \text{ccgt} \\
 \pi_c P_{c \rightarrow c} \pi_c P_{c \rightarrow c} \pi_a P_{a \rightarrow g} \pi_t P_{t \rightarrow t} \\
 = 0.4 \times 0.983 \times 0.4 \times 0.983 \times 0.1 \times 0.007 \times 0.3 \times 0.979 \\
 = 0.0000300
 \end{array}$$

Verosimilitud de ir desde la primera a la segunda secuencia = 0.0000300

En caso de que haya diferentes longitudes de ramas:

- para las *ramas muy cortas*, la probabilidad de que un carácter permanezca inmutable es alta, y la probabilidad de cambio es baja.
- para las *ramas largas*, aumenta la probabilidad de cambio de caracteres y se reduce la probabilidad de mantener estados.

Esto genera el problema de **long branch attraction**, que ocurre cuando grupos que han evolucionado rápidamente son colocados erróneamente en la base de los árboles filogenéticos al contar con más cambios



en sus secuencias.

La

atracción de ramas largas es un fenómeno que se produce cuando se infiere que los linajes que evolucionan rápidamente están estrechamente relacionados, independientemente de sus verdaderas relaciones evolutivas.

Inferencia Bayesiana (BI)

La inferencia bayesiana se basa en la **probabilidad posterior** de un árbol condicionada por la matriz observada (alineamiento de secuencias): la probabilidad de que el árbol sea correcto dados los datos.

La máxima verosimilitud (ML) examina modelos cuyos parámetros son constantes, y obtiene la verosimilitud de obtener los datos dados esos parámetros. BI también usa la verosimilitud de los datos, pero usando modelos cuyos parámetros son variables aleatorias con distribuciones estadísticas.

Antes del análisis de los datos, se asigna una distribución inicial a los parámetros del modelo (prior), que combinada con la verosimilitud de los datos permite calcular la probabilidad posterior de dichos parámetros.

El teorema de Bayes sirve para calcular esa probabilidad:

$$\Pr(H|D) = \frac{\Pr(D|H) \times \Pr(H)}{\Pr(D)}$$

Probabilidad de los datos dado el árbol (*likelihood*)

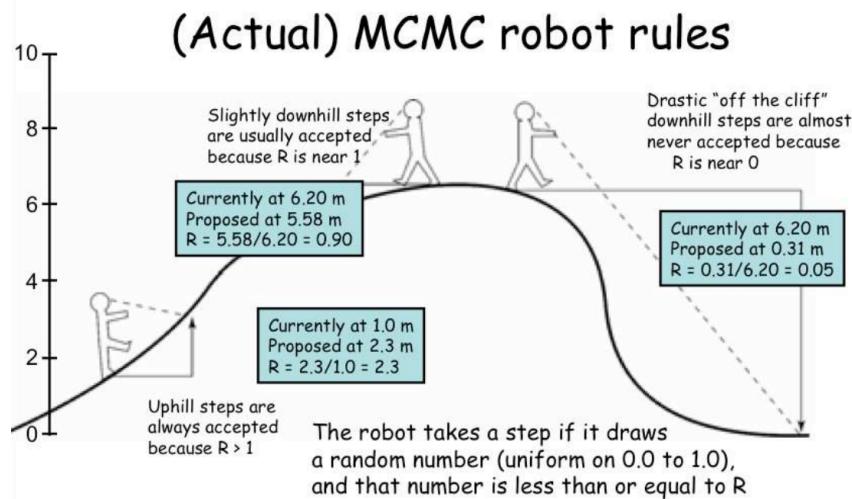
Probabilidad inicial del árbol (*prior*)

Probabilidad del árbol dados los datos

Probabilidad de los datos a través de todos los árboles

¿Cuál es la probabilidad inicial de un árbol? No se puede calcular el denominador de la expresión.

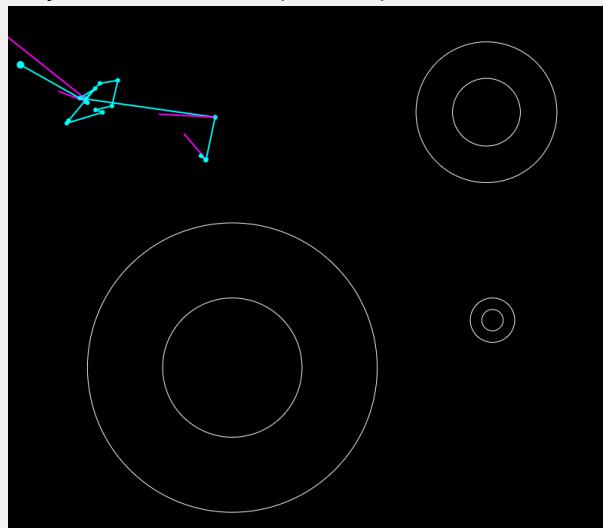
Sin embargo, sí se puede aproximar la probabilidad posterior usando un procedimiento de muestreo que se acelera mediante simulaciones de **Monte Carlo con cadenas de Markov** (MCMC).



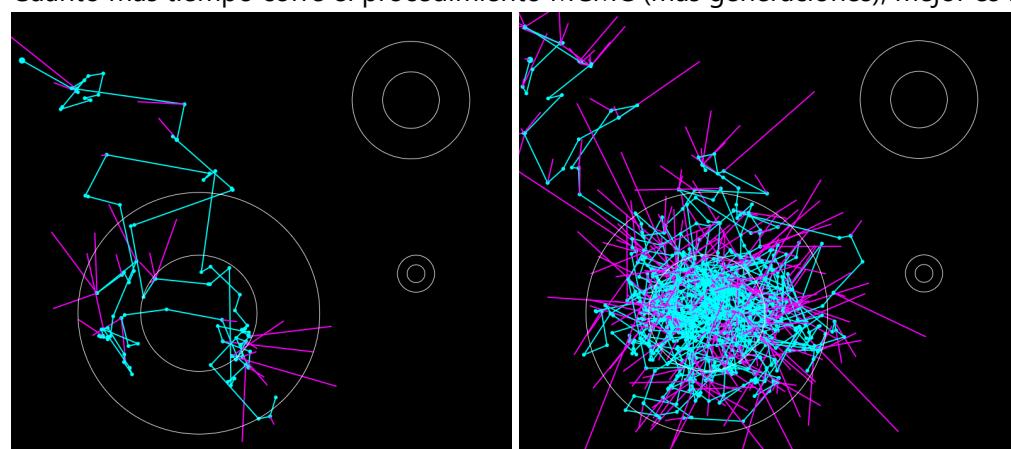
La idea es vagar al azar en el espacio de árboles de manera que se genera una distribución de árboles cuya media es la de la distribución deseada (la probabilidad Bayesiana).

Así, la inferencia bayesiana utiliza MCMC como herramienta para explorar todos los árboles posibles. Hace una buena aproximación del paisaje tras el periodo de "burnin".

Como burnin se conocen las primeras búsquedas al estar más alejadas de los árboles más probables (hay un consenso de quitar el primer 20% de las búsquedas por este efecto).



Cuanto más tiempo corre el procedimiento MCMC (más generaciones), mejor es la aproximación.



100 y 1000 ciclos

Conviene repetir el proceso empezando desde diferentes puntos (random seeds).

Se realizan dos búsquedas:

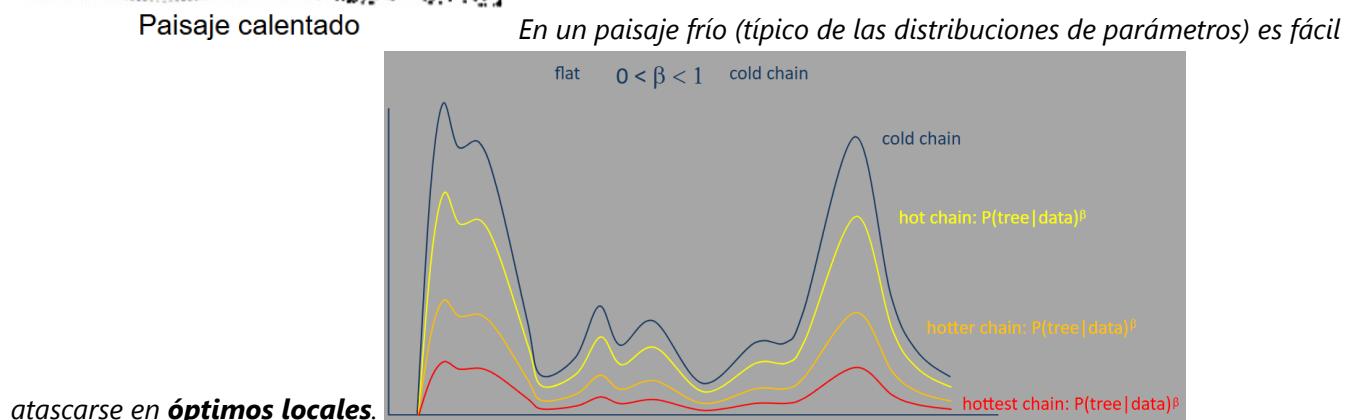
- Las cadenas frías contienen colinas altas y valles profundos.
- Las cadenas calientes son paisajes en los que la travesía entre colinas es más fácil.



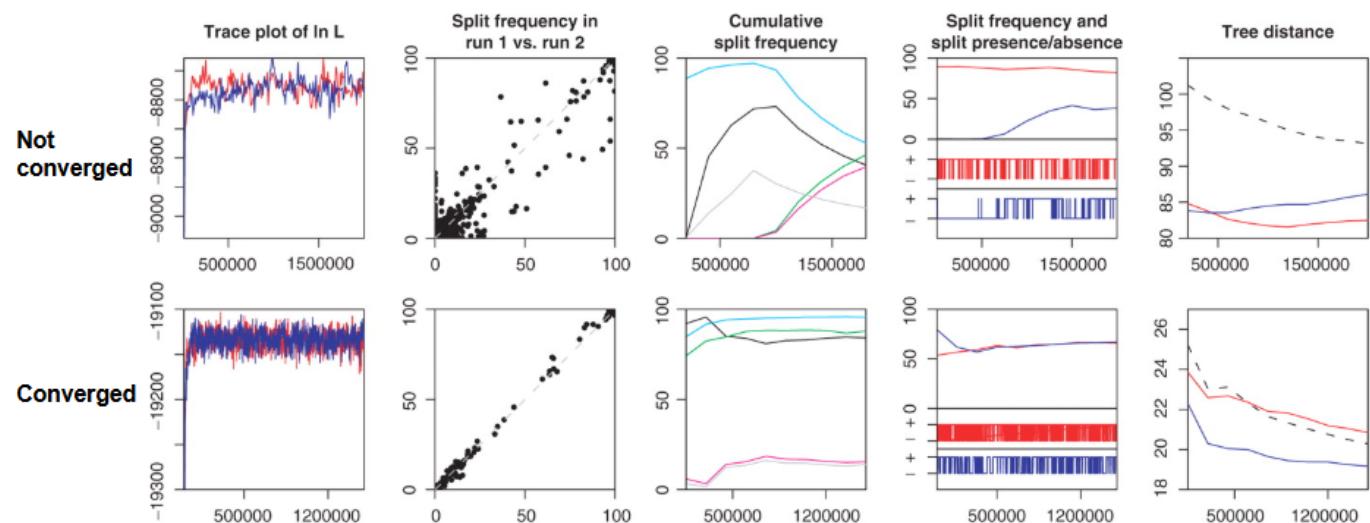
Paisaje frío

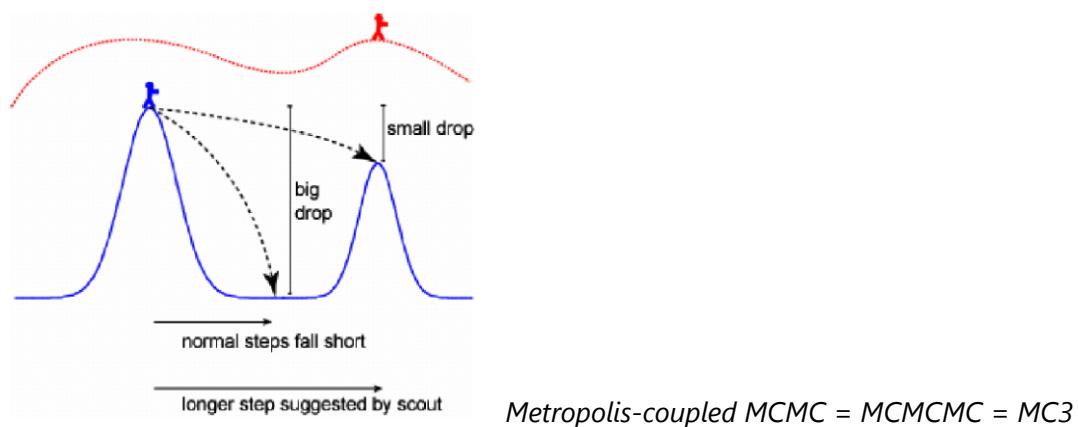


Paisaje calentado



Por ello, se utilizan técnicas de mezclado para resolver el problema de los óptimos locales: se busca que **las dos cadenas sean convergentes**.





Los **criterios de convergencia** son:

- Las distribuciones posteriores de los parámetros del modelo son similares entre cadenas independientes.
- Las probabilidades posteriores de los clados son similares entre cadenas independientes comenzadas con topologías aleatorias (desviación estándar promedio de las frecuencias divididas).

En la práctica ("resumen"):

1. Empezar el procedimiento de MCMC con un árbol aleatorio y un modelo con parámetros aleatorios.
2. En cada generación, y al azar, se propone:
 - Un nuevo árbol (que se acepta o se rechaza)
 - Un nuevo valor para los parámetros del modelo (que se acepta o se rechaza)
3. Se ejecutan en paralelo una cadena fría y varias calentadas (que orientan a la fría a través del espacio de árboles), mediante un procedimiento MCMCMC.
4. Cada k generaciones, se muestrea los valores de los parámetros de la cadena fría.
5. Tras n generaciones se obtiene una distribución muestral (la cadena fría habrá pasado más tiempo en los mejores lugares del espacio de árboles) y una desviación estándar de las frecuencias divididas.

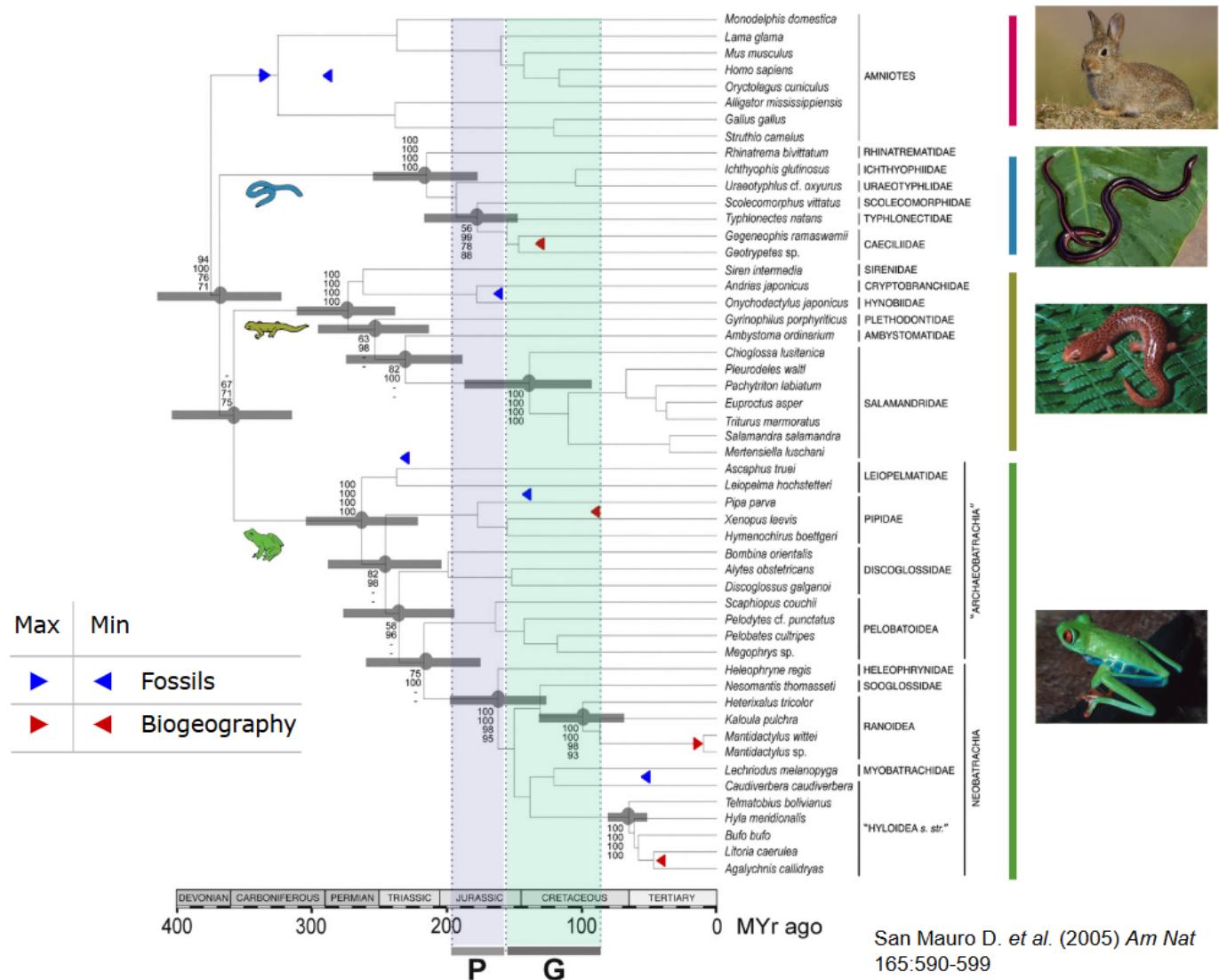
Interpretación y soporte

El método de inferencia Bayesiana calcula una **probabilidad posterior (BPP)** para cada nodo, que va del 0 al 1.

La interpretación estadística es inmediata: es la **probabilidad de que el clado sea cierto, dado un modelo, unas premisas y unos datos**.

Sin embargo, las BPP suelen ser *sospechosamente altas* (tendencia a la sobreestimación), mucho más que los valores de apoyo de bootstrap. Por ello, se suelen quedar con los valores bayesianos a partir de un 0.9, pero un bootstrap a partir de 70.

Strengths	Weaknesses
Parsimony methods	<ul style="list-style-type: none"> • Simplicity and intuitive appeal • The only framework appropriate for some data (such as SINES and LINES)
Distance methods	<ul style="list-style-type: none"> • Assumptions are implicit and poorly understood • Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution • Branch lengths are substantially underestimated when substitution rates are high • Maximum parsimony may suffer from long-branch attraction
Likelihood methods	
<ul style="list-style-type: none"> • Fast computational speed • Can be applied to any type of data as long as a genetic distance can be defined • Models for distance calculation can be chosen to fit data 	<ul style="list-style-type: none"> • Most distance methods, such as neighbour joining, do not consider variances of distance estimates • Distance calculation is problematic when sequences are divergent and involve many alignment gaps • Negative branch lengths are not meaningful
Bayesian methods	
<ul style="list-style-type: none"> • Can use complex substitution models to approach biological reality • Powerful framework for estimating parameters and testing hypotheses 	<ul style="list-style-type: none"> • Maximum likelihood iteration involves heavy computation • The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret

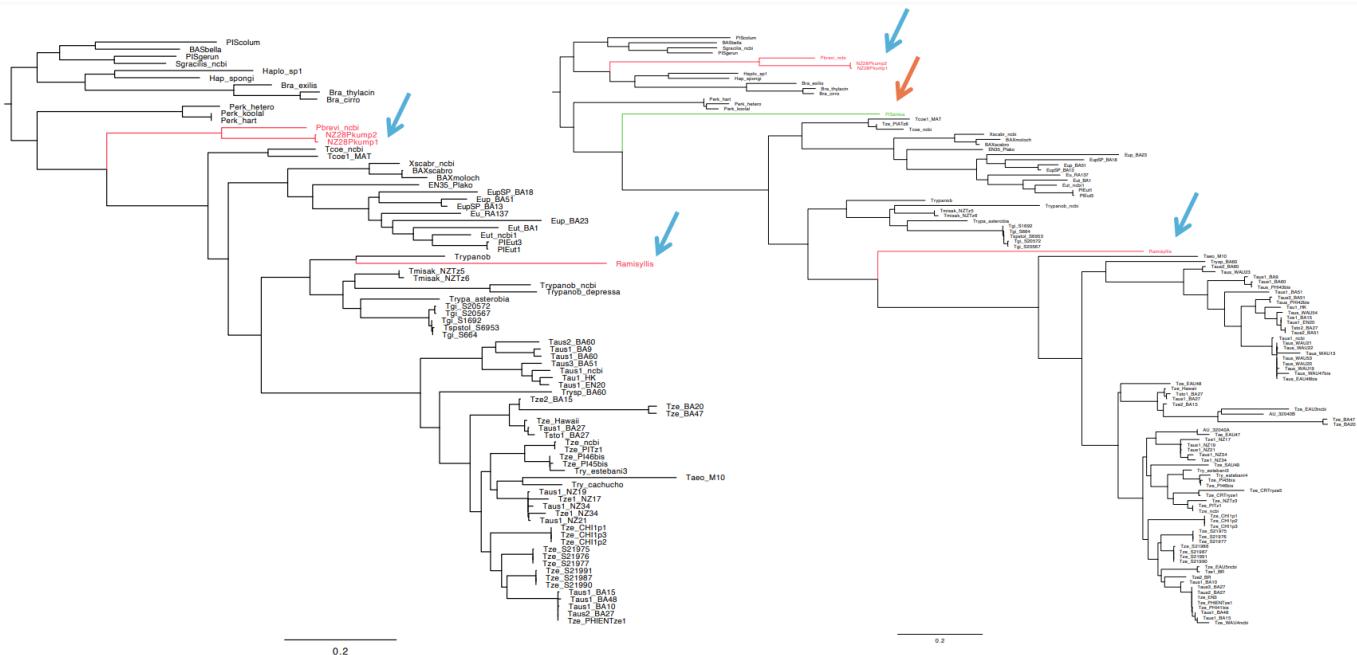


Ejemplo de cronograma. La datación se ha basado en registros fósiles (flechas azules) y eventos biogeográficos (flechas rojas).

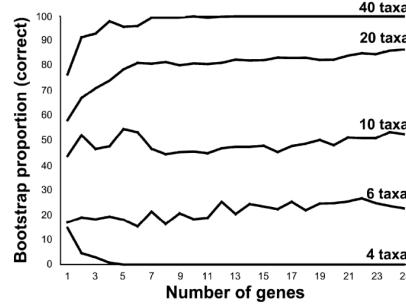
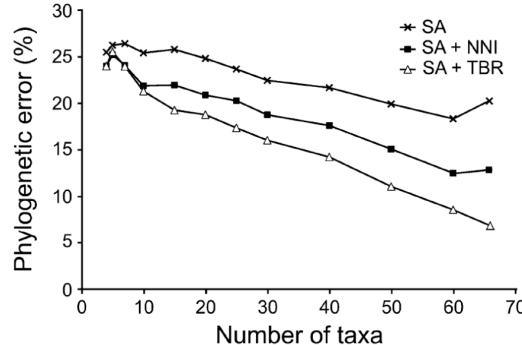
What do we need to build a phylogenetic tree?

1. What **taxa**?

- Para construir un árbol filogenético, se necesitan elegir los taxones. Esto es decisión nuestra, pero hay que intentar no tener sesgos y elegir un buen ingroup y outgroup.
- Impact of the **outgroup** taxa selection:



- Impact of the **ingroup** taxa selection:



2. Data / Markers:

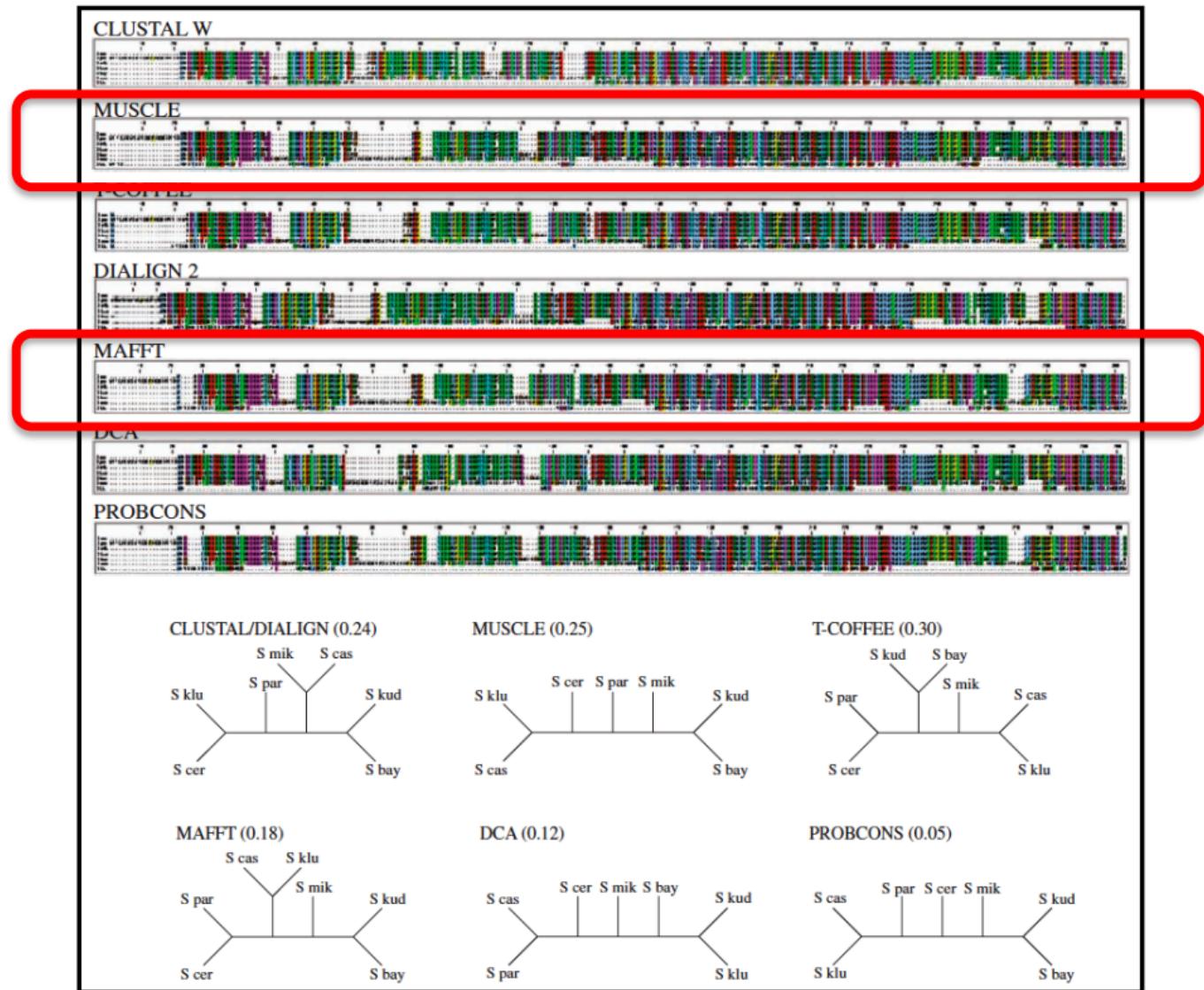
- What type of **data**? - Your decision, but based on what is available!
- How many **markers**? - depende del presupuesto que haya... pero se debe intentar tener los máximos posibles por PCR u otros.

Variability vs. Conservation: Depending on your question, but most (small scale) phylogenetic studies now in *metazoans* use:

- Ribosomal genes (18S, 28S, 16S, ITS1&2) which are more conserved
- Protein coding genes (COI, COB, ALG11, etc) which are more variable

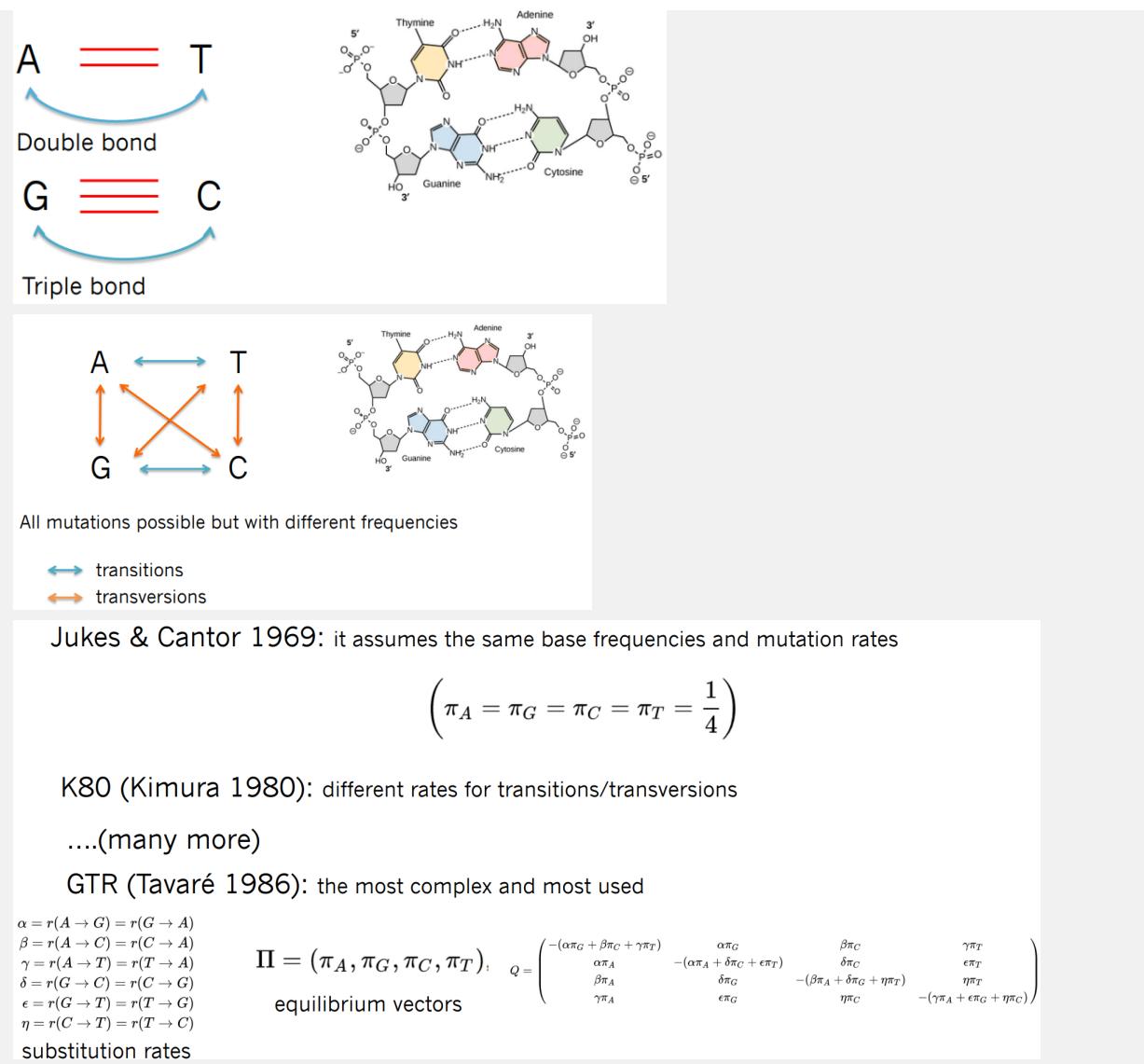
Do your research... and check ncbi first to see what is available already

3. What **alignment program**? - Muscle? MAFFT?



4. What **model of evolution**? - jModelTest.

- Not all genes evolve (mutate) at the same pace and in the same manner.
- Their mutation rates depend on many parameters:
 - GC genome content
 - Genome size
 - Generation time
 - Expression levels (usage of the gene)
 - Protein coding genes
 - Position in the genome
 - Linkage disequilibrium
 - etc
- Therefore, each gene will have its own substitution model. These substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution.



- **jModelTest** is a tool to carry out statistical selection of best-fit models of nucleotide substitution. It implements five different model selection strategies: hierarchical and dynamical likelihood ratio tests (hLRT and dLRT), Akaike and Bayesian information criteria (AIC and BIC), and a decision theory method (DT).

5. What **phylogenetic method**?

Finalmente se debe elegir un método filogenético (por ejemplo, máxima verosimilitud + inferencia bayesiana).

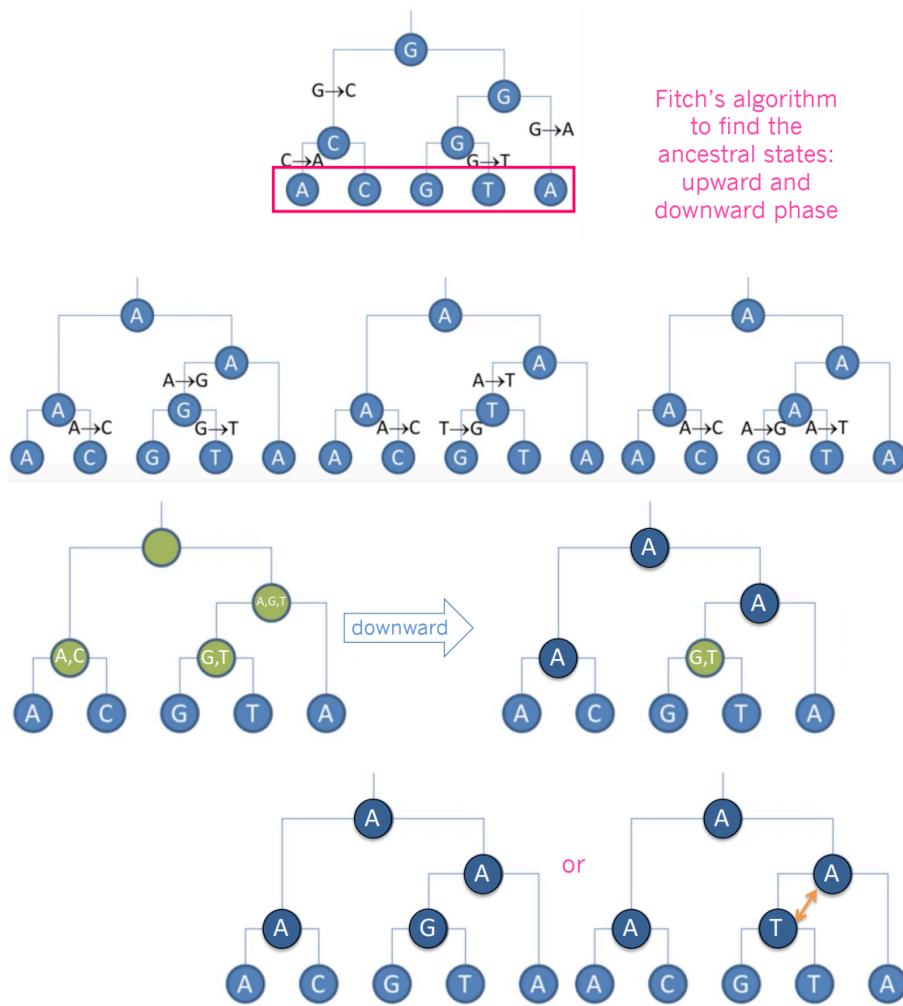
MAXIMUM PARSIMONY

- The assumption is that the 'true' tree will contain the least number of mutations possible i.e., the most parsimonious solution.
- Given a set of sequences (partial evidence) we need to find ancestral sequences, build a rooted tree, and estimate the smallest number of changes contained in the branches
- It poses an impossible computational problem (no algorithm known) and therefore all models use simplified versions...

Weighted parsimony: different scoring for each change (transition/transversion) and different scoring for each position:

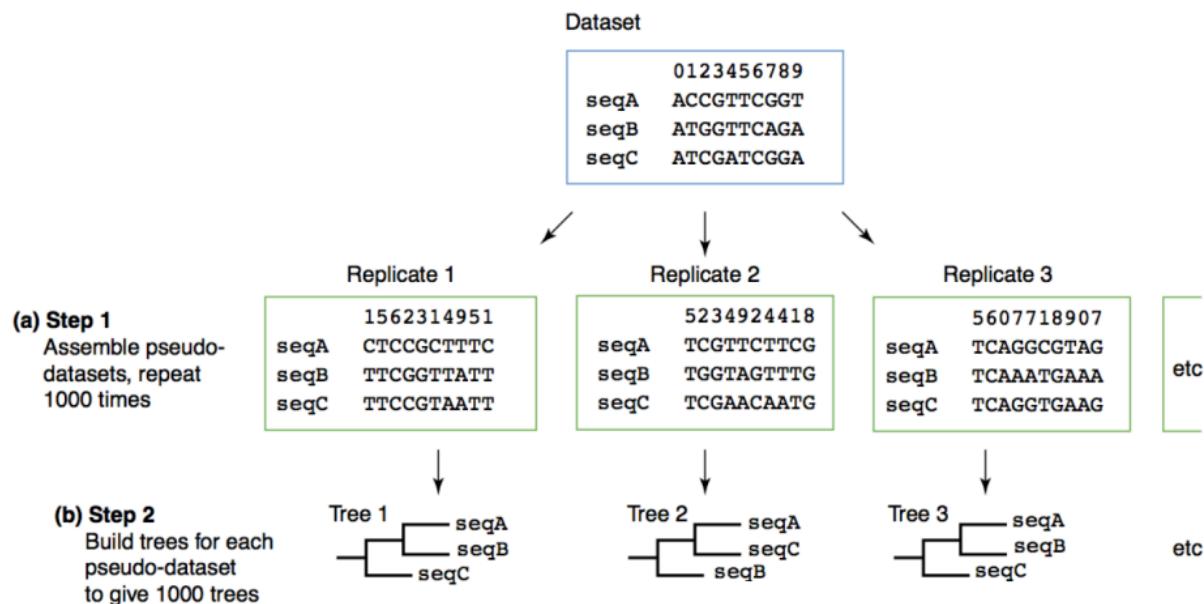
	1	2	3	4	5	6	7	8	9	10	
Species 1 - A	G	G	G	T	A	A	C	T	G		
Species 2 - A	C	G	A	T	T	A	T	T	A		
Species 3 - A	T	A	A	T	T	G	T	C	T		
Species 4 - A	A	T	G	T	T	G	T	C	G		

Informative sites



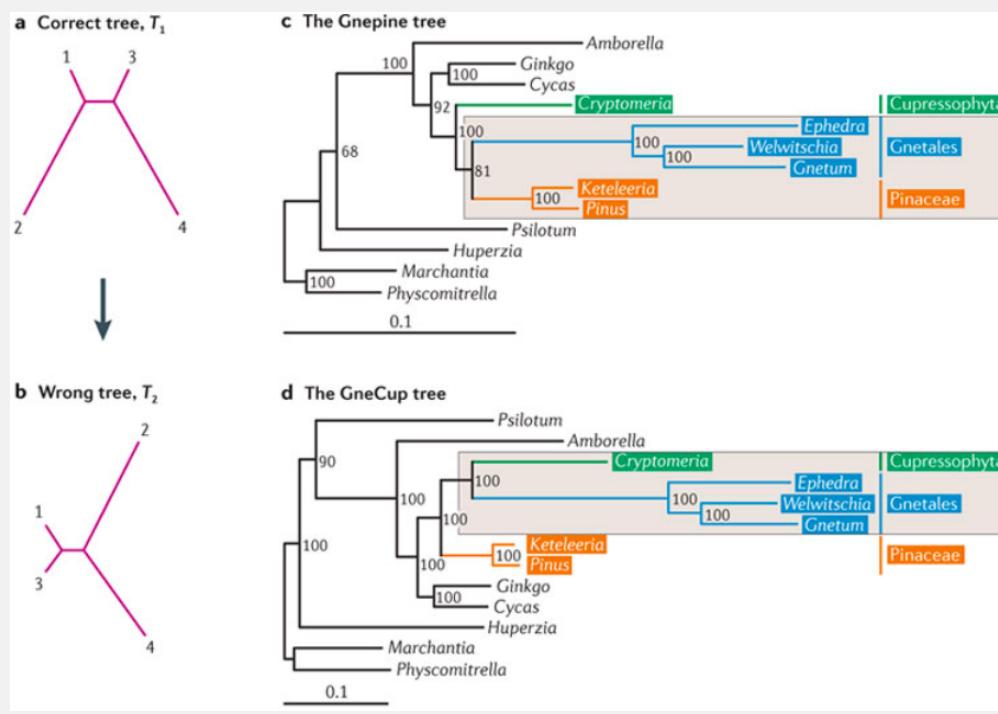
MAXIMUM LIKELIHOOD

- Likelihood: the probability of producing the observed data by a model given the model parameters,
 $LD = \Pr(X|\Theta)$
 - X: data already aligned, each site mutates independently
 - Θ : parameters of the model (topology, model of substitution)
- In big datasets the computation is extremely difficult and long, BUT it is one of the most powerful methods because:
 - it uses models of substitution (evolutionary models)
 - it corrects multiple substitutions
 - it allows estimation of branch length (=amount of change from ancestor)
 - fixing all but one parameter it allows finding the max. likelihood for that parameter



Problems with maximum likelihood (and parsimony)

Long branch attraction: is a phenomenon when **rapidly evolving lineages** are inferred to be closely related, regardless of their true evolutionary relationships



BAYESIAN INFERENCE

Based on Bayes' theorem, the bayesian approach combines the prior probability of a tree $P(A)$ with the likelihood of the data (B) to produce a posterior probability distribution on trees $P(A|B)$.

Likelihood gives you the probability of the data given the hypothesis and *Bayesian gives you the probability of the hypothesis given the data.*



Posterior probabilities

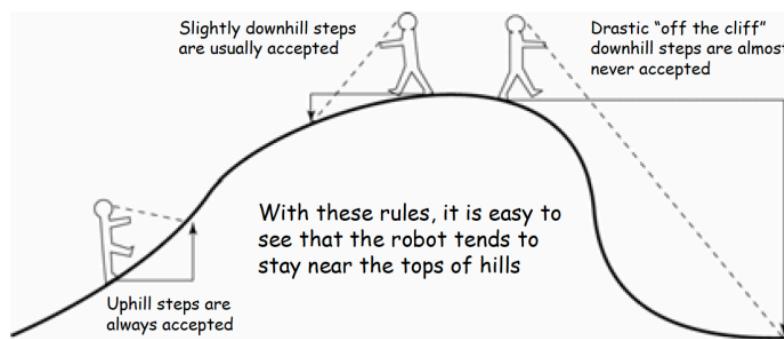
↓

It uses MCMC (Markov chain-Montecarlo) algorithms

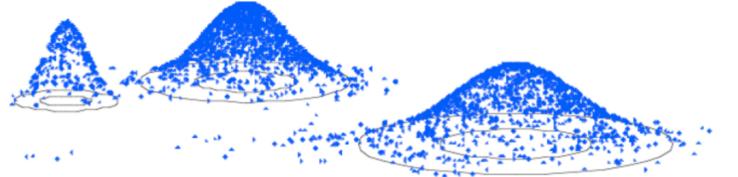
- **How does it work?**

- Start somewhere (that somewhere will have a likelihood and a prior).
- Will move randomly and propose a new state (maybe adjust one branch length), if the change has a better *likelihood x prior*, the chain goes there.
- Calculate the posterior probability ratio between the current and previous state. It should be between 0 and 1.

MCMC robot's rules



- Choose a random number between 0 and 1 and if that number has a better likelihood than the ratio of states, the change is accepted (sometimes if slightly worse as well).
- That is how the chain crosses likelihood valleys.



6. How to **assess confidence**?

There are multiple ways for **evaluating support**:

- **Bremer support [> 70 %]:**

- Only in parsimony (MP)
- Difference in the branch lengths when clades are removed

- **Jackknife [> 70 %]:**

- Parsimony (MP)
- probability of a clade observed in all the trees

- **Bootstrap [> 70 %]:**

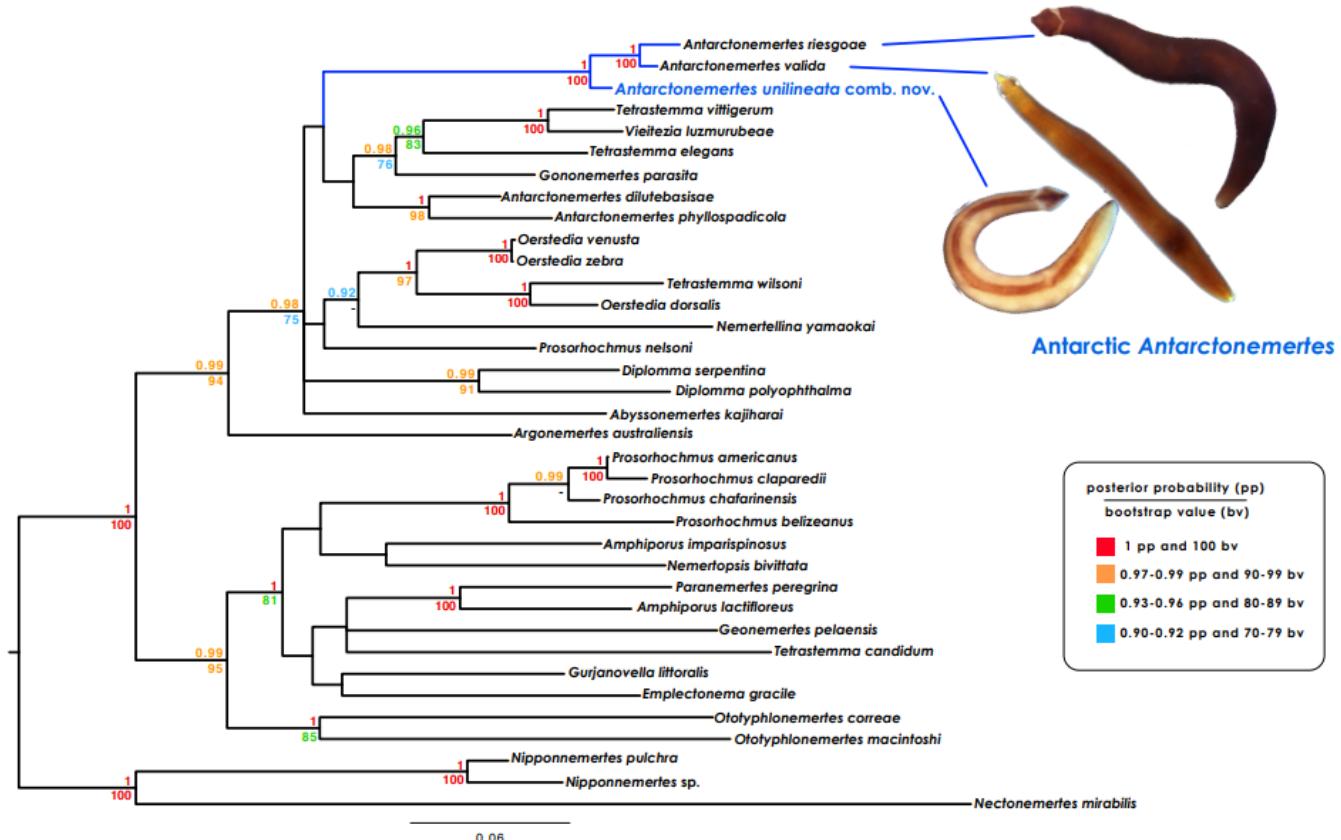
- Máxima verosimilitud (ML)
- probability of a clade observed in all the trees

- **Posterior probability [> 0.95]:**

- inferencia bayesiana (BI)
- probability of a clade being assigned under the conditions sampled

- **Convergence:**

- inferencia bayesiana (BI)
- assessing whether all chains (MCMC) converged in the same solution



Ejercicios

Ejercicio 1: Construir un árbol a partir de un alineamiento.

Dado el siguiente alineamiento, construye un árbol filogenético.

	1	2	3	4
A	a	a	a	b
B	a	a	a	b
C	b	a	a	b
D	b	b	a	a
E	b	b	a	a
F	b	b	b	b

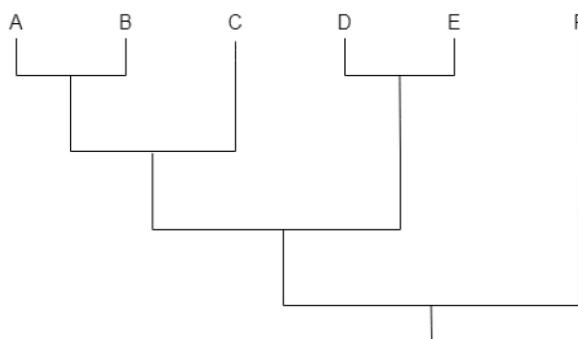
Empezar por aquellos alineamientos que son iguales:

- Los taxones A y B son iguales, al igual que D y E. Por tanto, sabemos que esos taxones van juntos (fechas azul y rojas).
- Despues, se busca el taxón más similar. En este caso, el taxón C es el siguiente más similar a los taxones A y B.

	1	2	3	4	
A	a	a	a	b	A y B tienen 3 posiciones compartidas con C
B	a	a	a	b	
C	b	a	a	b	
D	b	b	a	a	
E	b	b	a	a	D y E solo comparten 2 posiciones con C
F	b	b	b	b	

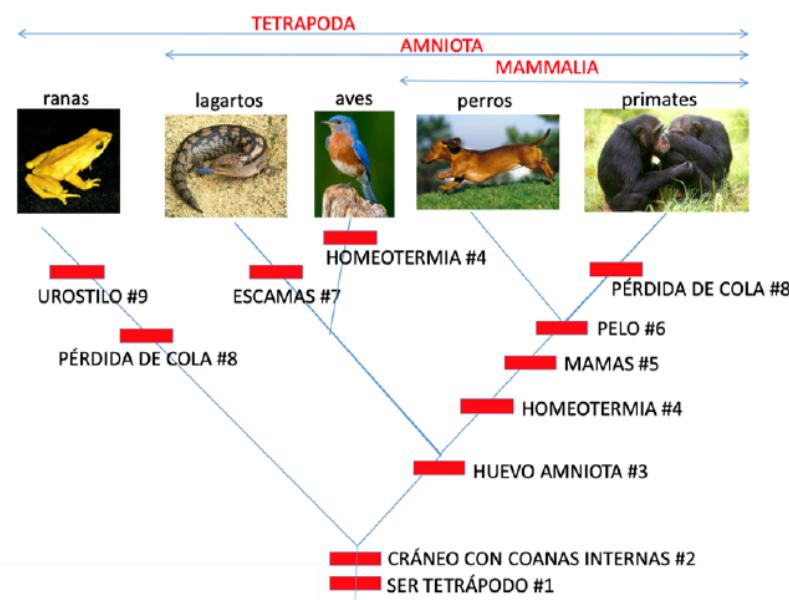
- Por último, el taxón F es el más diferente, siendo por tanto el outgroup.

El árbol quedaría así:



Ejercicio 2: Construir la matriz de caracteres desde un árbol.

Dado el siguiente árbol filogenético con las distintas mutaciones, construye una matriz de alineamiento.



Pasos:

- Crear una tabla con las especies y las mutaciones en filas y columnas.
- Ir anotando si la mutación está presente (a) o ausente (b)
[mejor usar 0 (ausente) y 1 (presente)]

	1	2	3	4	5	6	7	8	9
Ranas	a	a	b	b	b	b	a	a	
Lagartos	a	a	a	b	b	b	a	b	a
Aves	a	a	a	a	b	b	b	b	
Perros	a	a	a	a	a	a	b	b	b
Primates	a	a	a	a	a	a	b	a	b

Ejercicio 3: Identificación de caracteres

En relación con el árbol filogenético del ejercicio 2, explica los caracteres en relación con todos los integrantes del árbol.

- Los caracteres 1 y 2 son **plesiomorfías** al ser ancestrales.
- El carácter 3 es una **sinapomorfía** (sería una **simplesiomorfía** para los amniotas, pero como nos piden relacionarlo con todos los taxones, es una sinapomorfía del árbol).
- El carácter 9 es una **autapomorfía** de las ranas, y el 7 de los lagartos.
- El carácter 8 es otra *autapomorfía con convergencia* entre ranas y primates, generando así un grupo **polifilético** (el ancestro común más cercano no tiene el carácter).
- El grupo hermano de los amniotas son las ranas.
- Así, el grupo hermano de los mamíferos son lagartos y aves; no son los amniotas porque los mamíferos están englobados en los amniotas.