

Introducción al análisis filogenético

Pablo Fresia

Unidad Mixta Pasteur + INIA

Unidad de Bioinformática

Institut Pasteur de Montevideo

Laboratorio de Genómica Evolutiva

Facultad de Ciencias, Udelar

16 de octubre de 2025



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



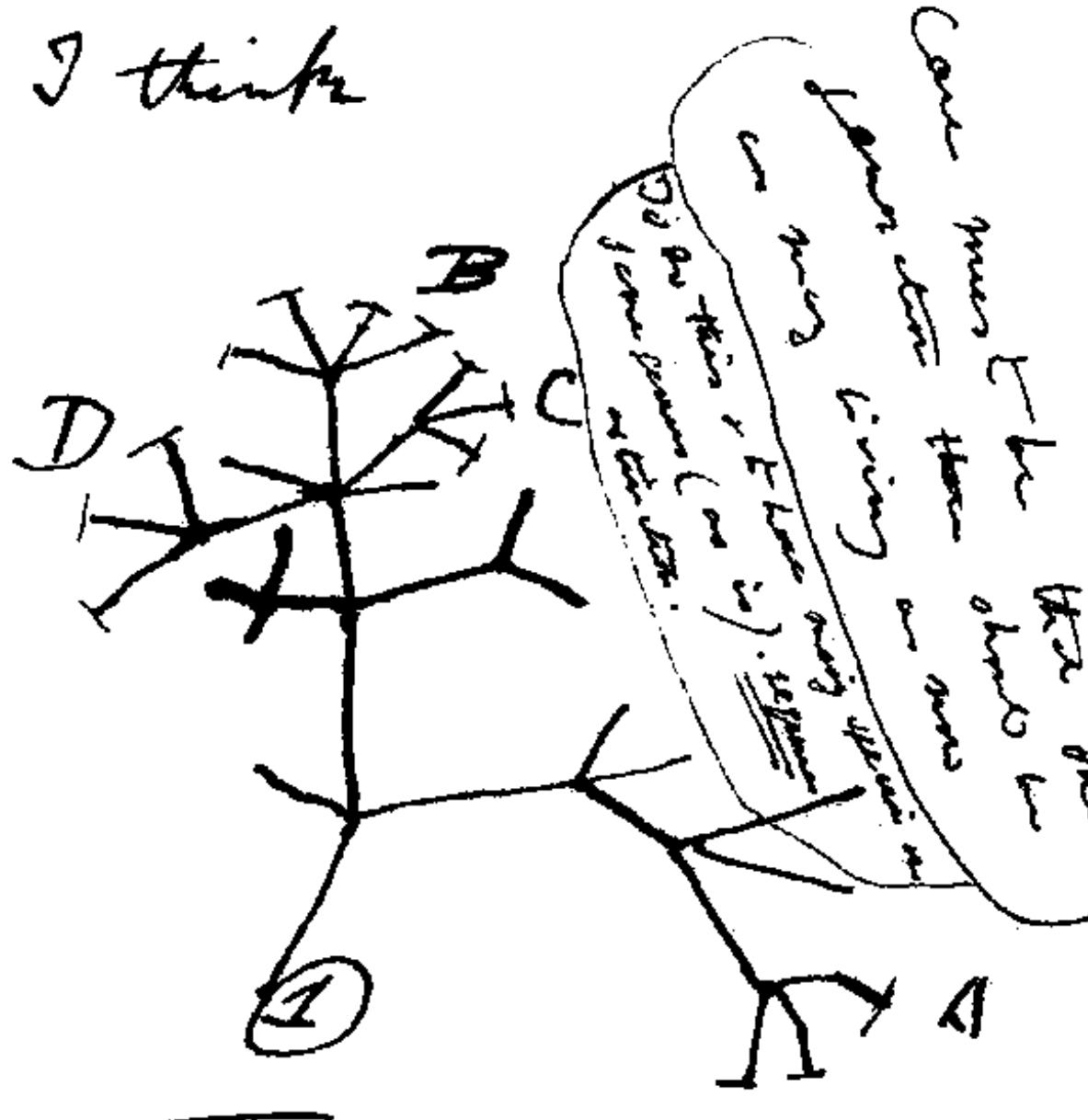
CURSO

Fundamentos y herramientas bioinformáticas para análisis genómicos, 13 al 17 de octubre de 2025

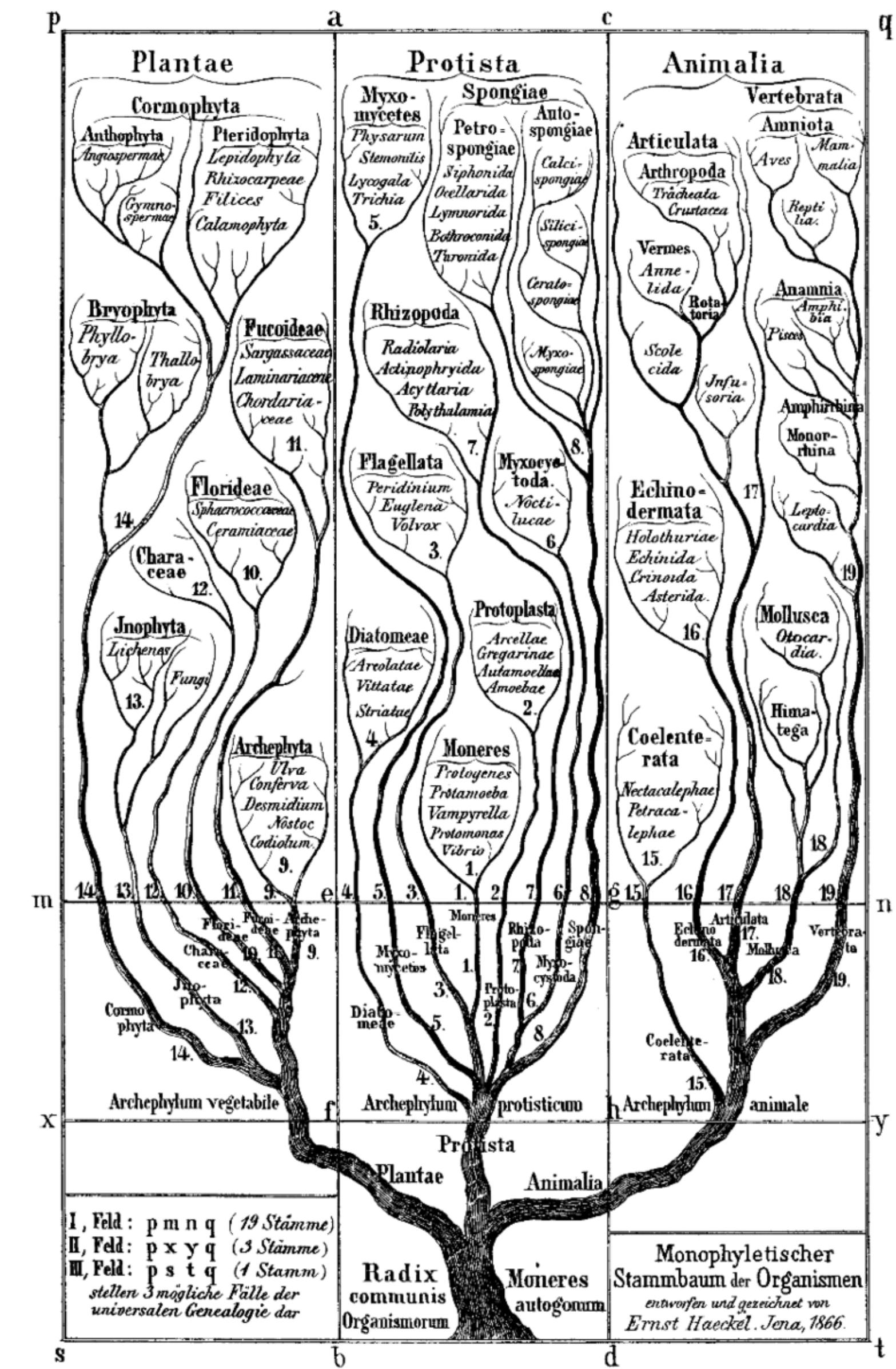
- ¿Que es una filogenia?
- Árboles filogenéticos
 - Terminología y representación
- Homología y analogía (homoplasia)
- Alineamiento y limpieza de alineamientos de ADN/Proteínas
- Modelos evolutivos
- Inferencia de una filogenia
 - Métodos por distancia
 - Máxima Verosimilitud
- Soporte estadístico del árbol
 - Bootstrap

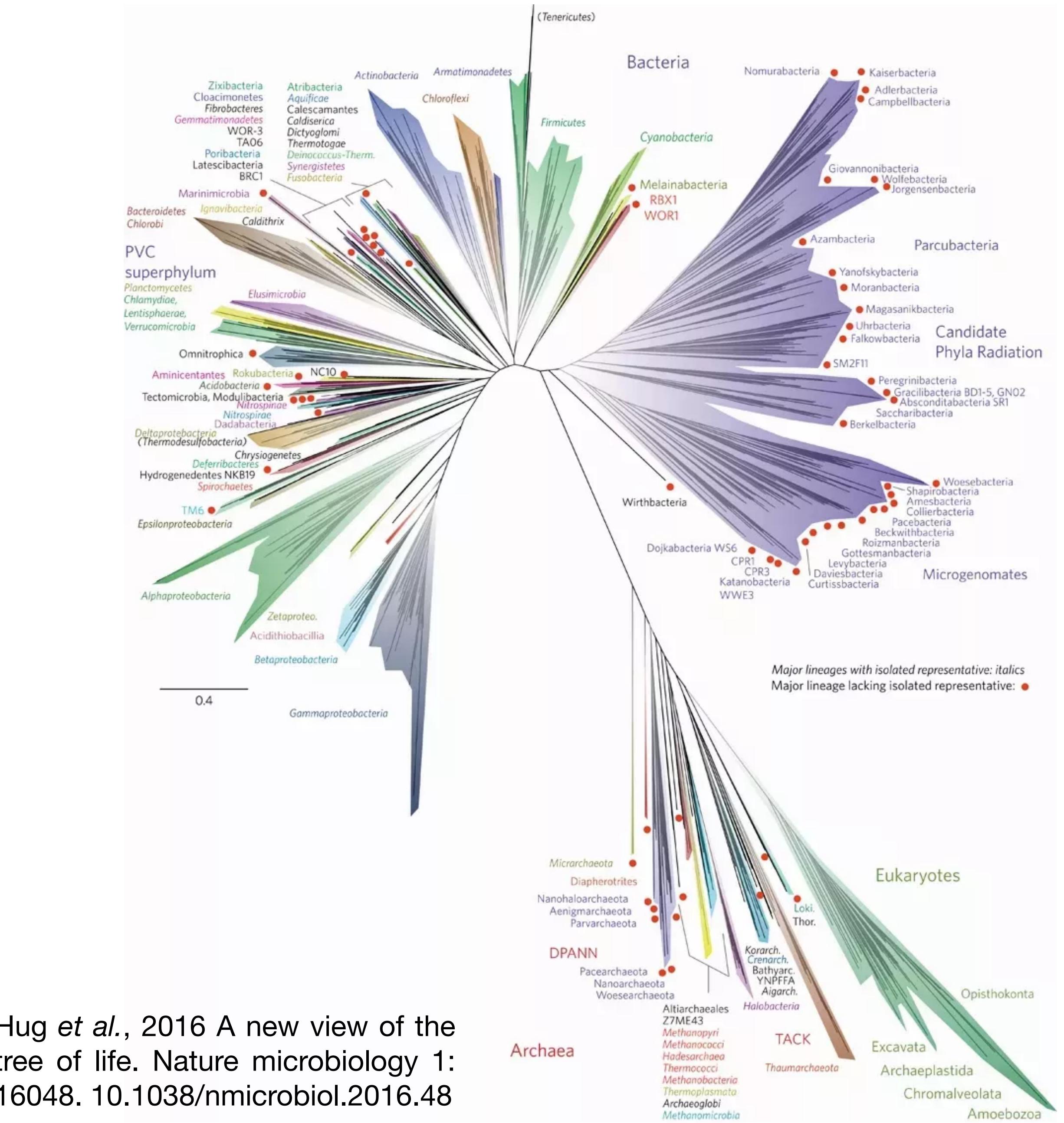
I think

5

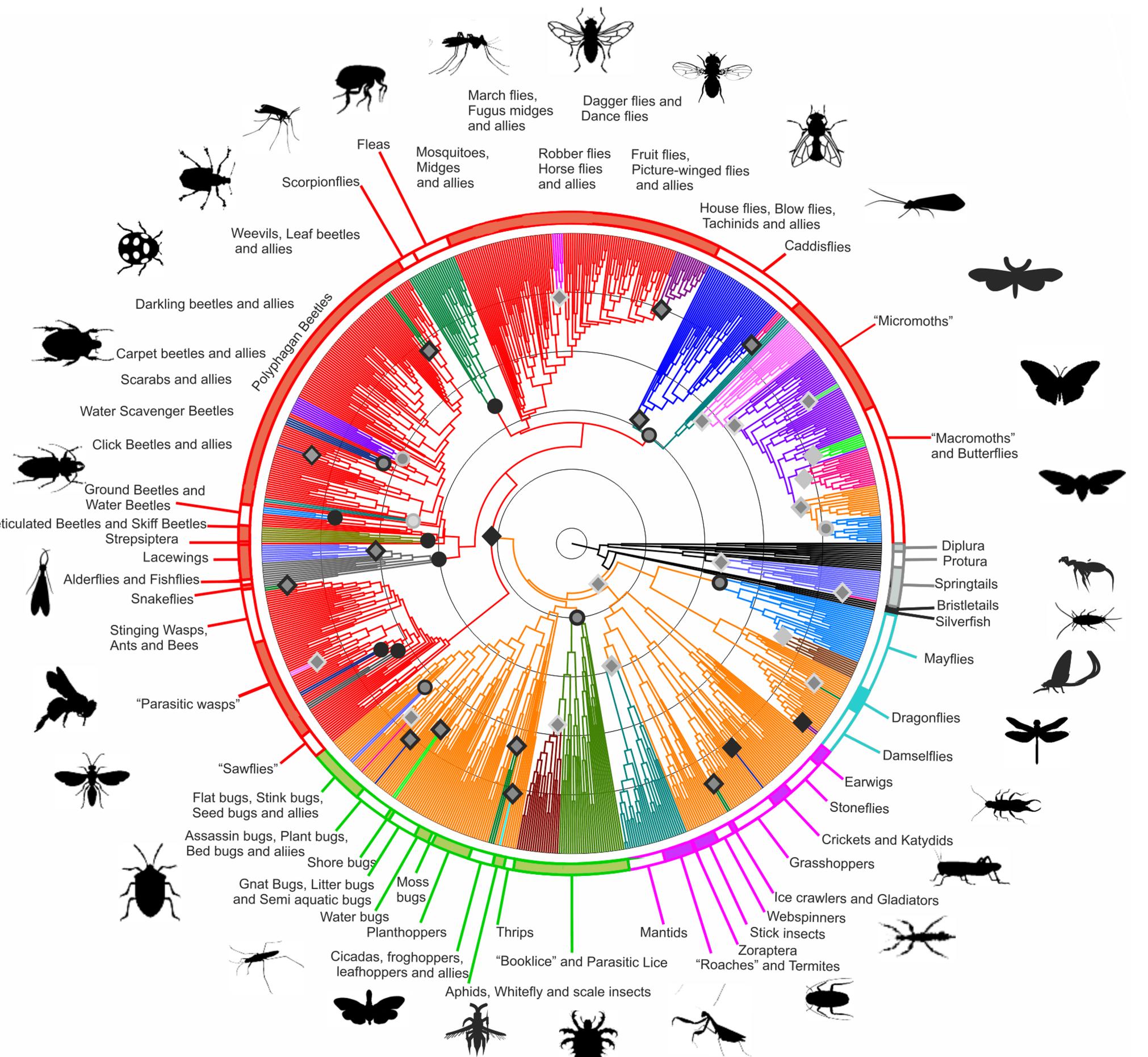


Then between A + B. various
ways of relation. C + B. The
finest gradation, B + D
rather greater distinction.
These forms would be
formed. - binary relation

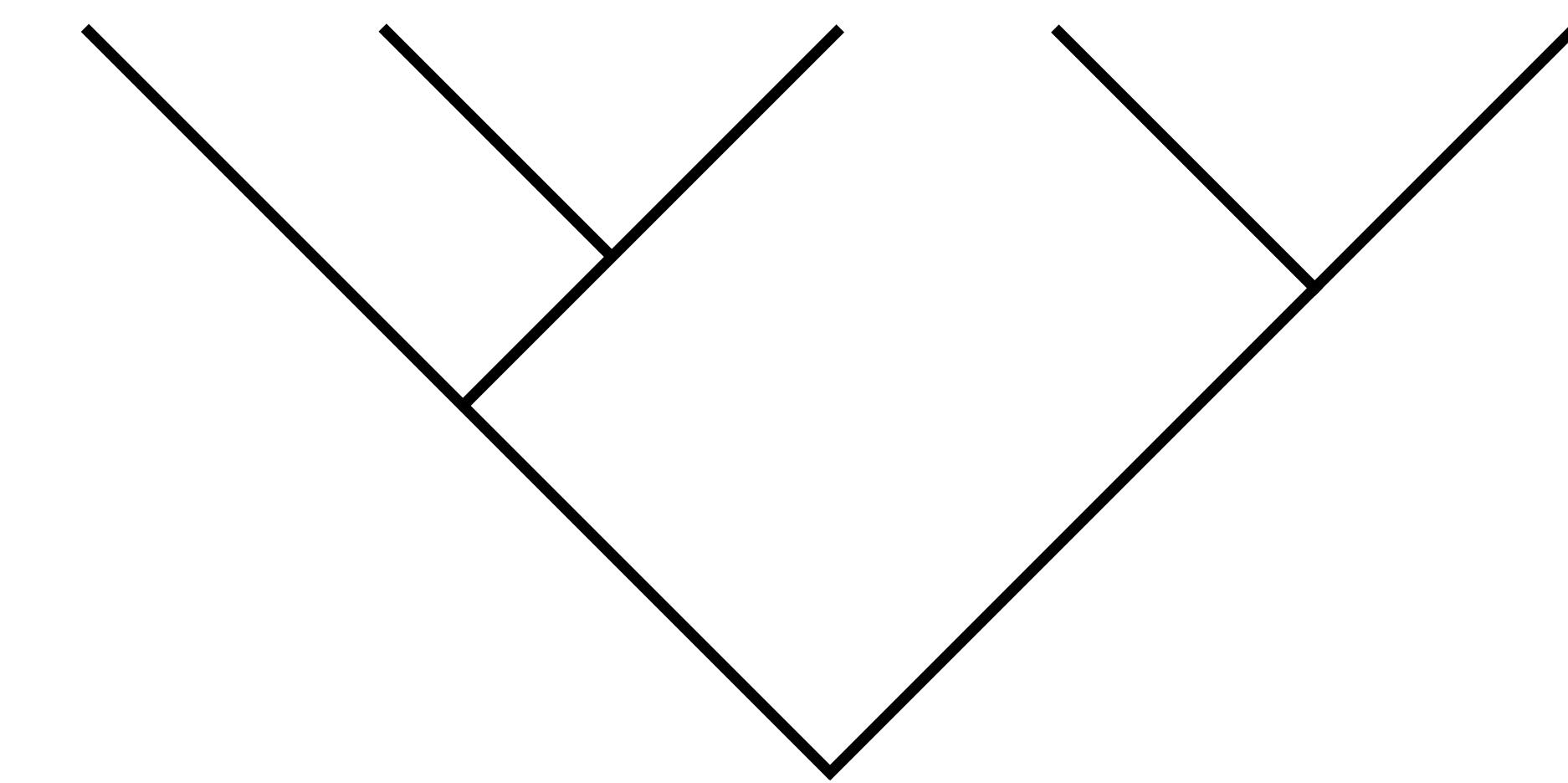




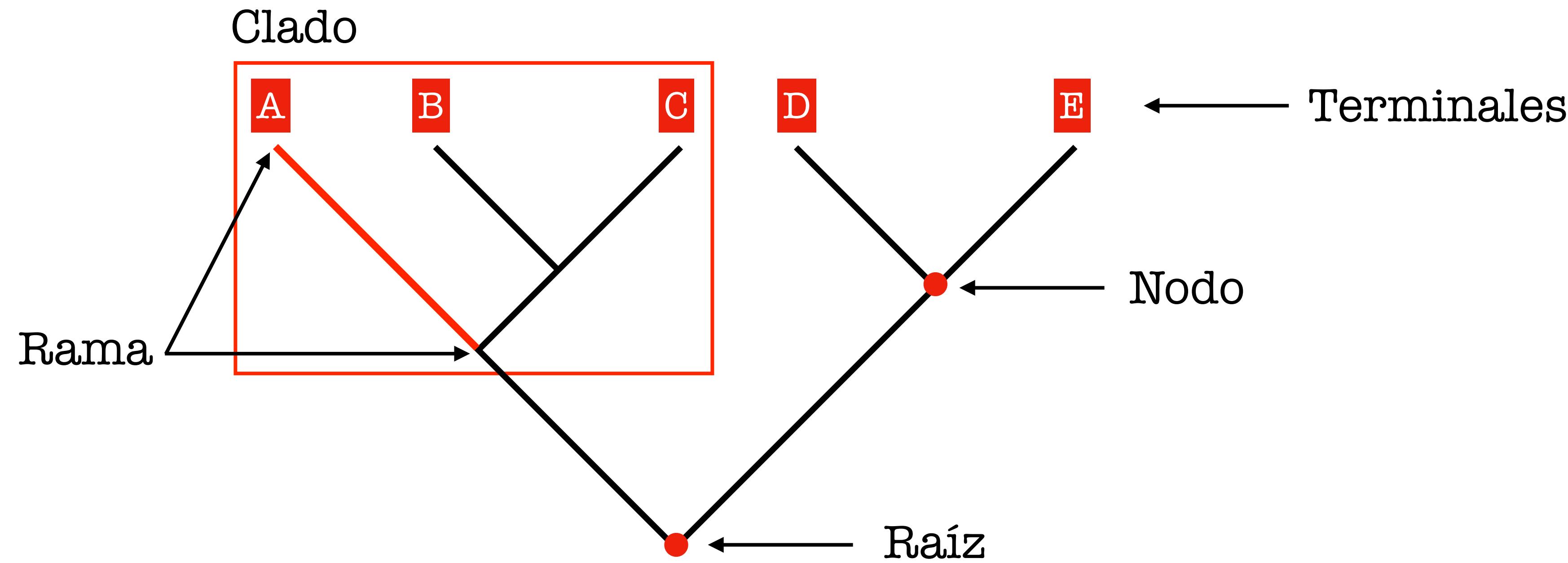
Filogenia de las familias de hexápodos actuales.
Rainford et al. 2014 PloS one 9: e109085



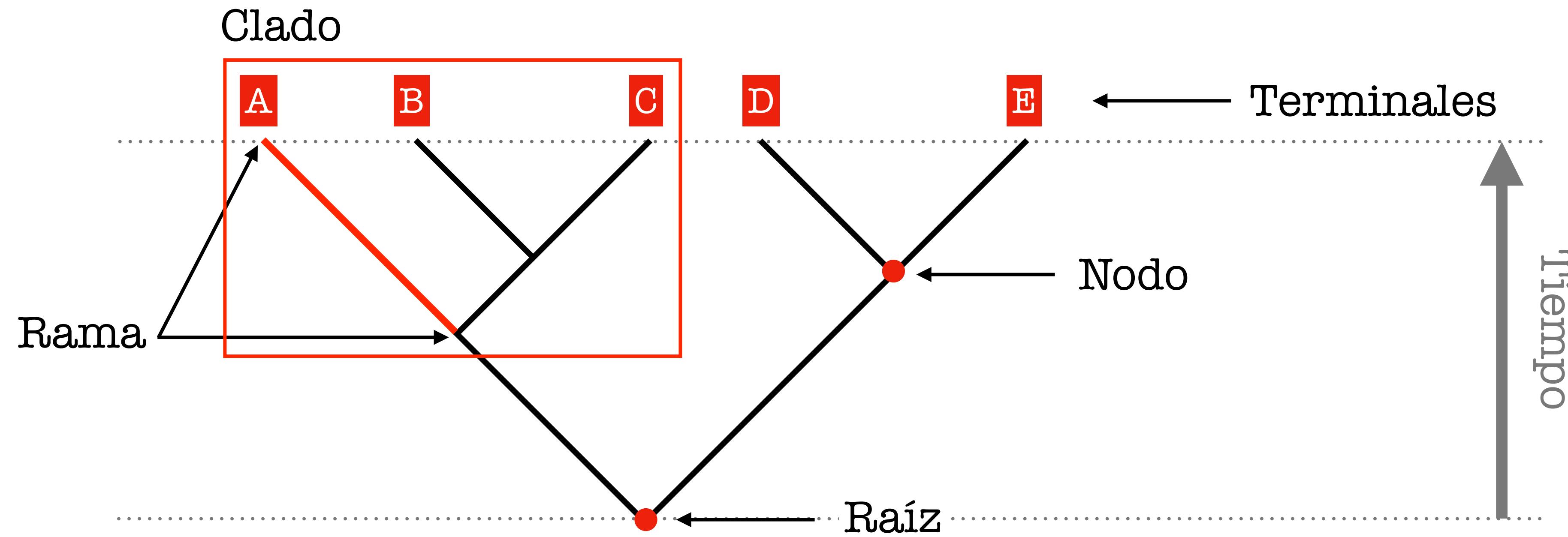
Árboles: terminología



Árboles: terminología

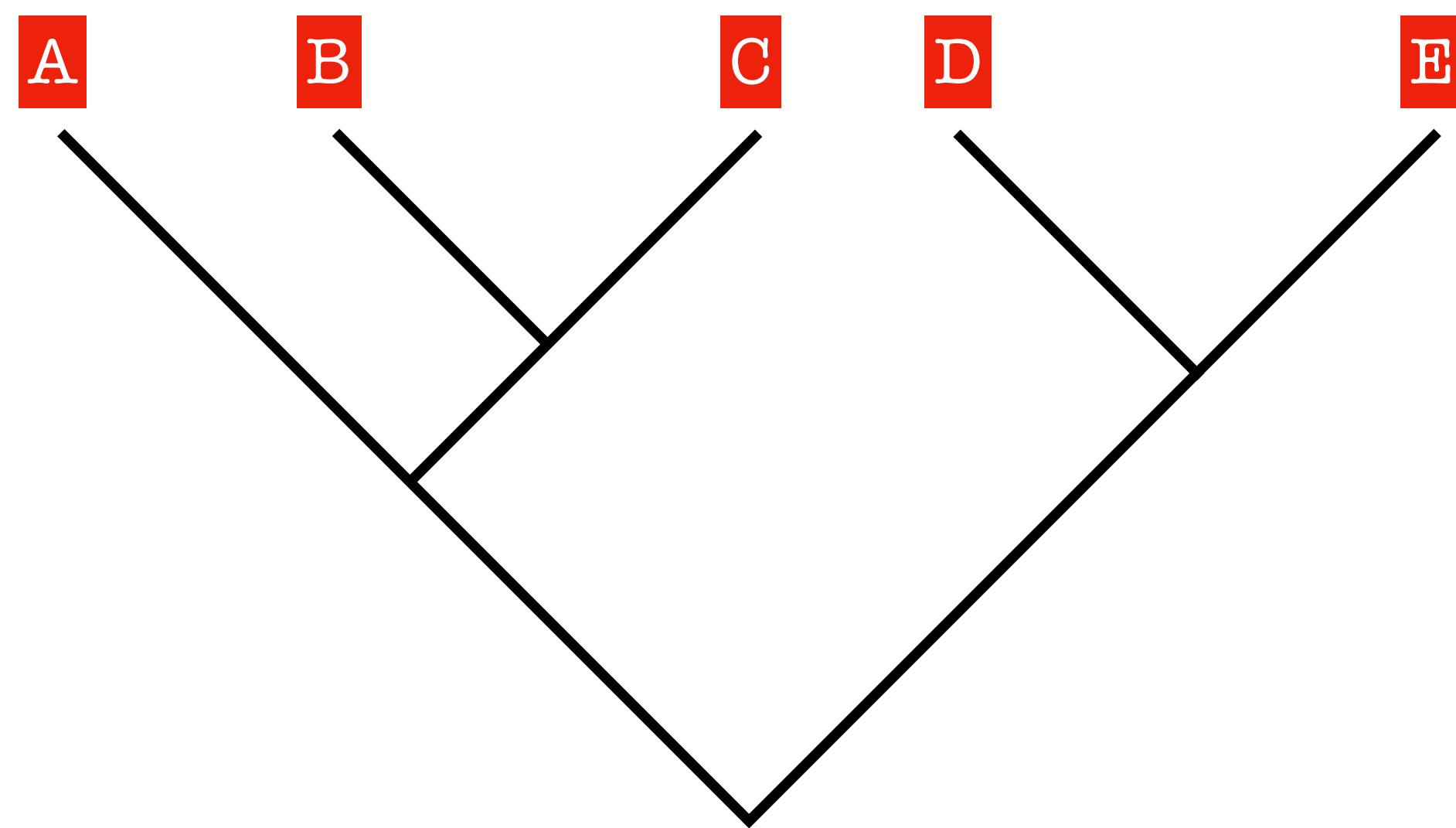


Árboles: terminología

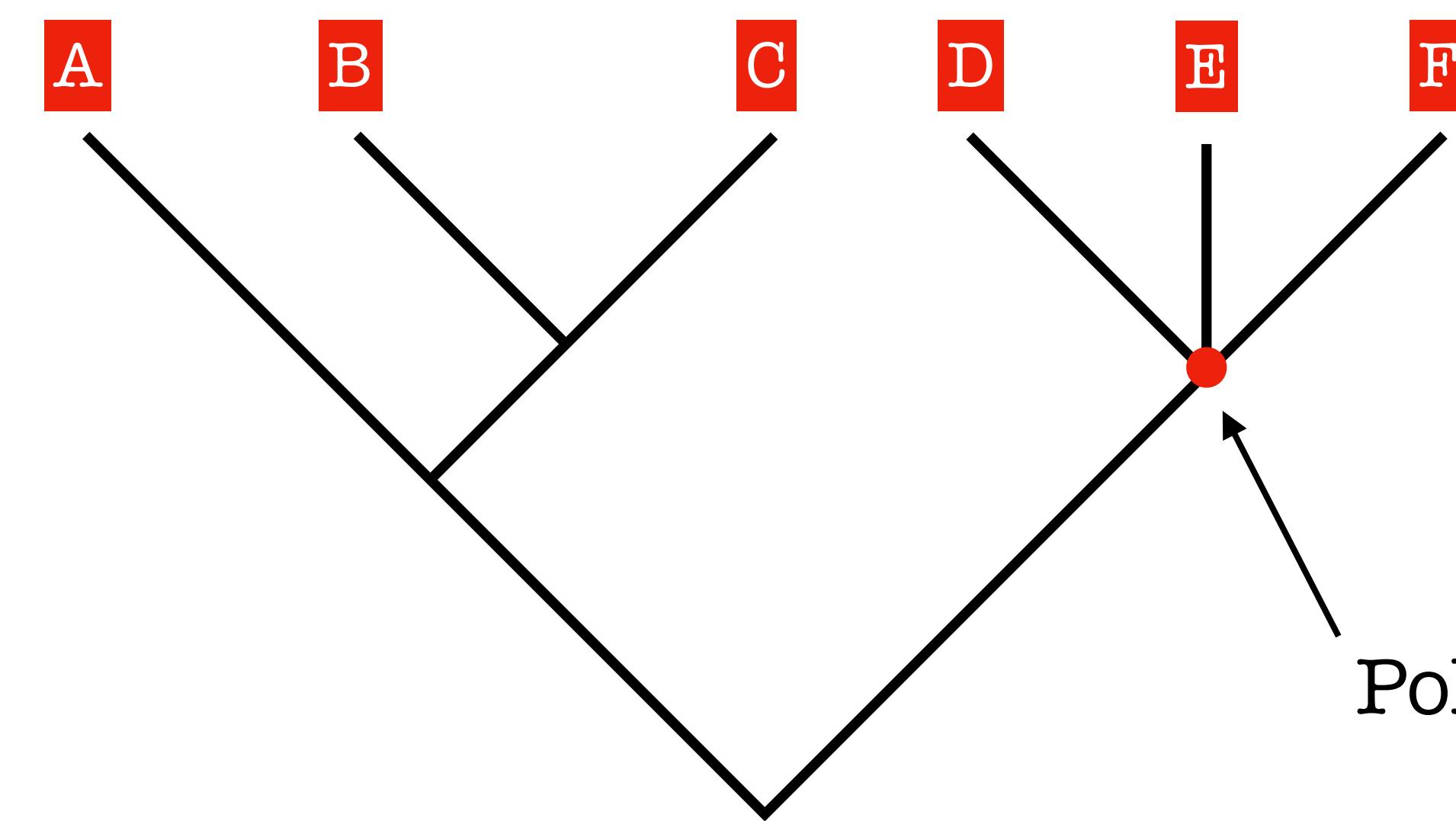


Formato Newick: (((B,C),A),(D,E));

Árboles: terminología



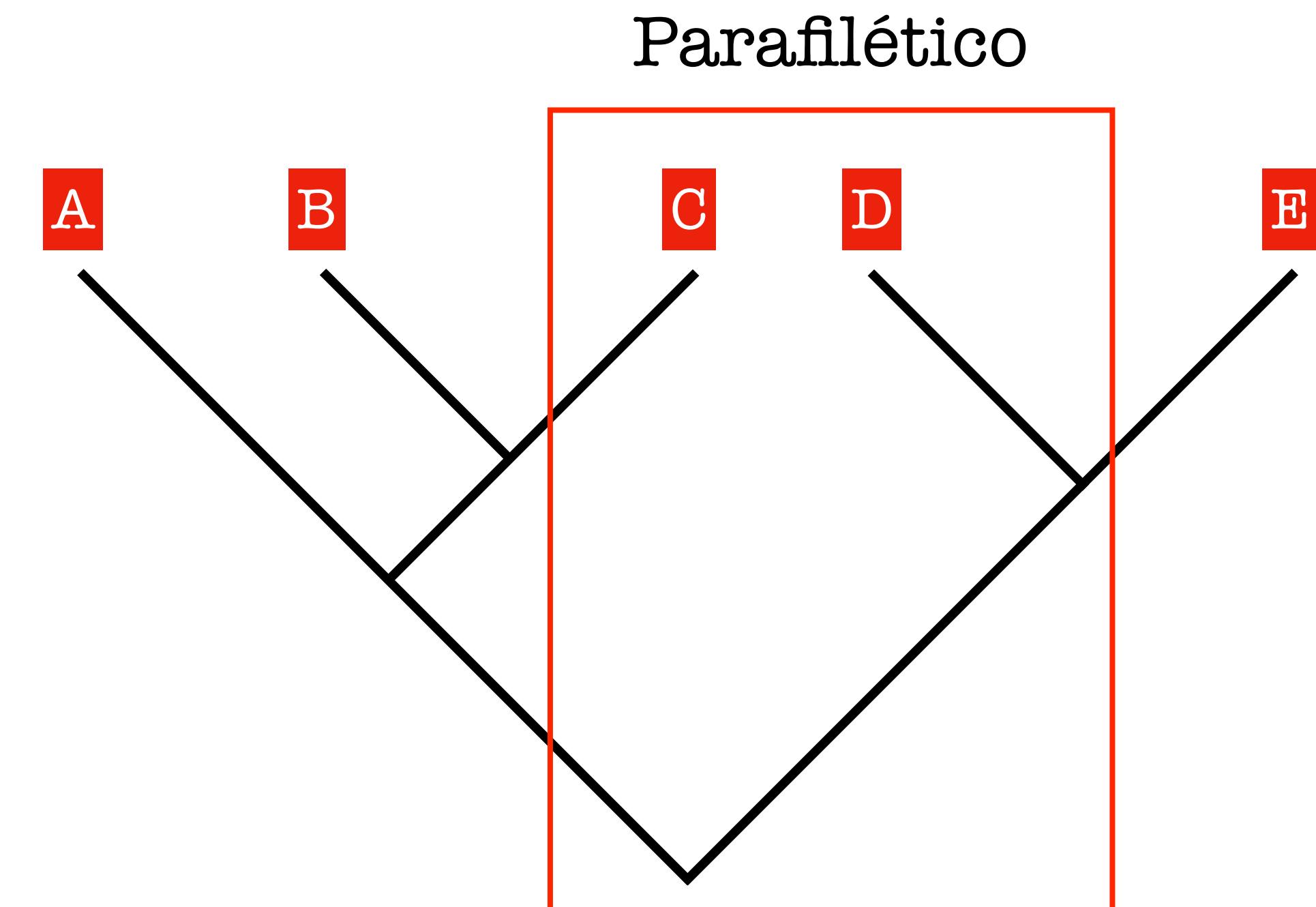
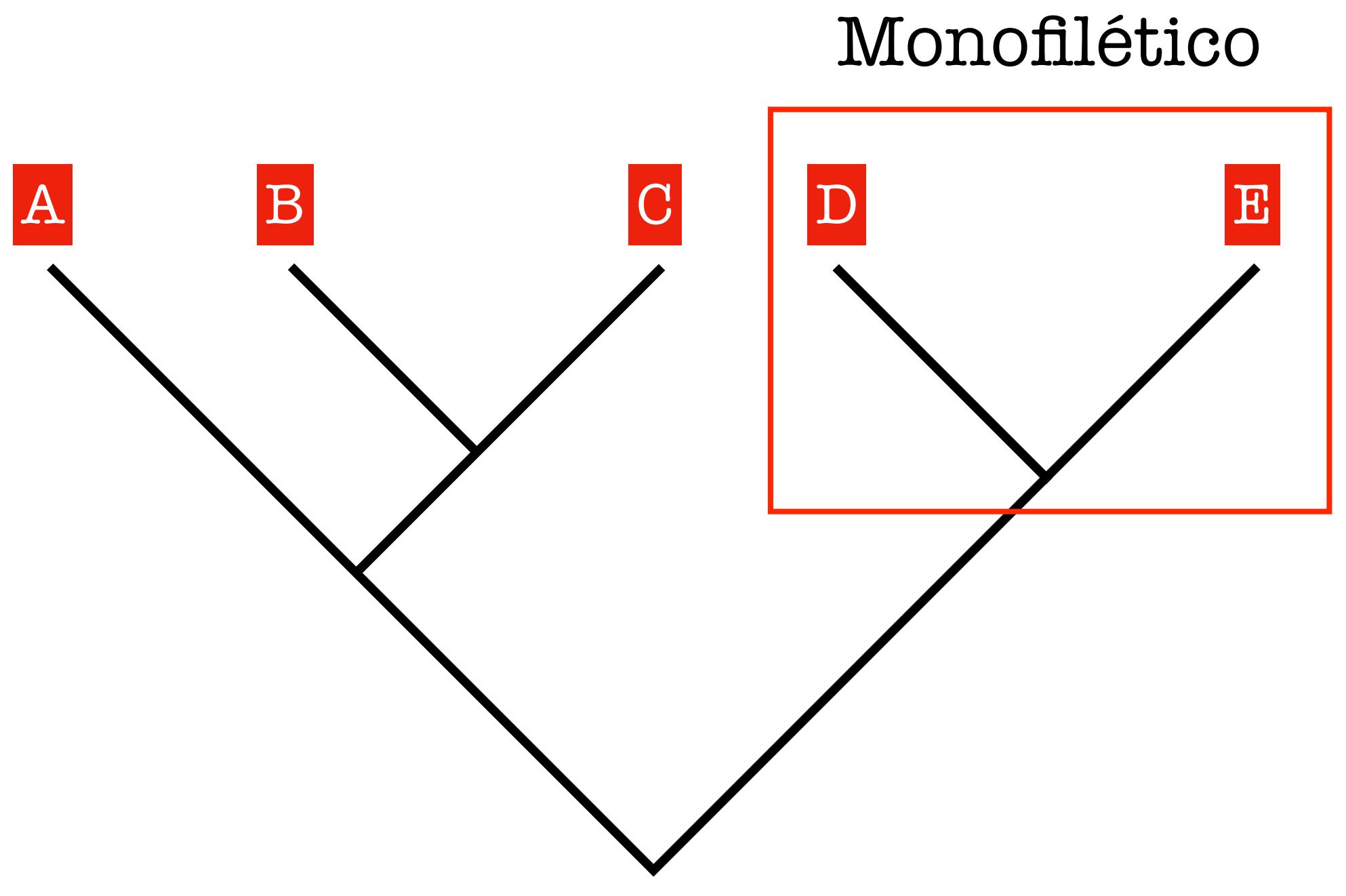
Totalmente resuelto



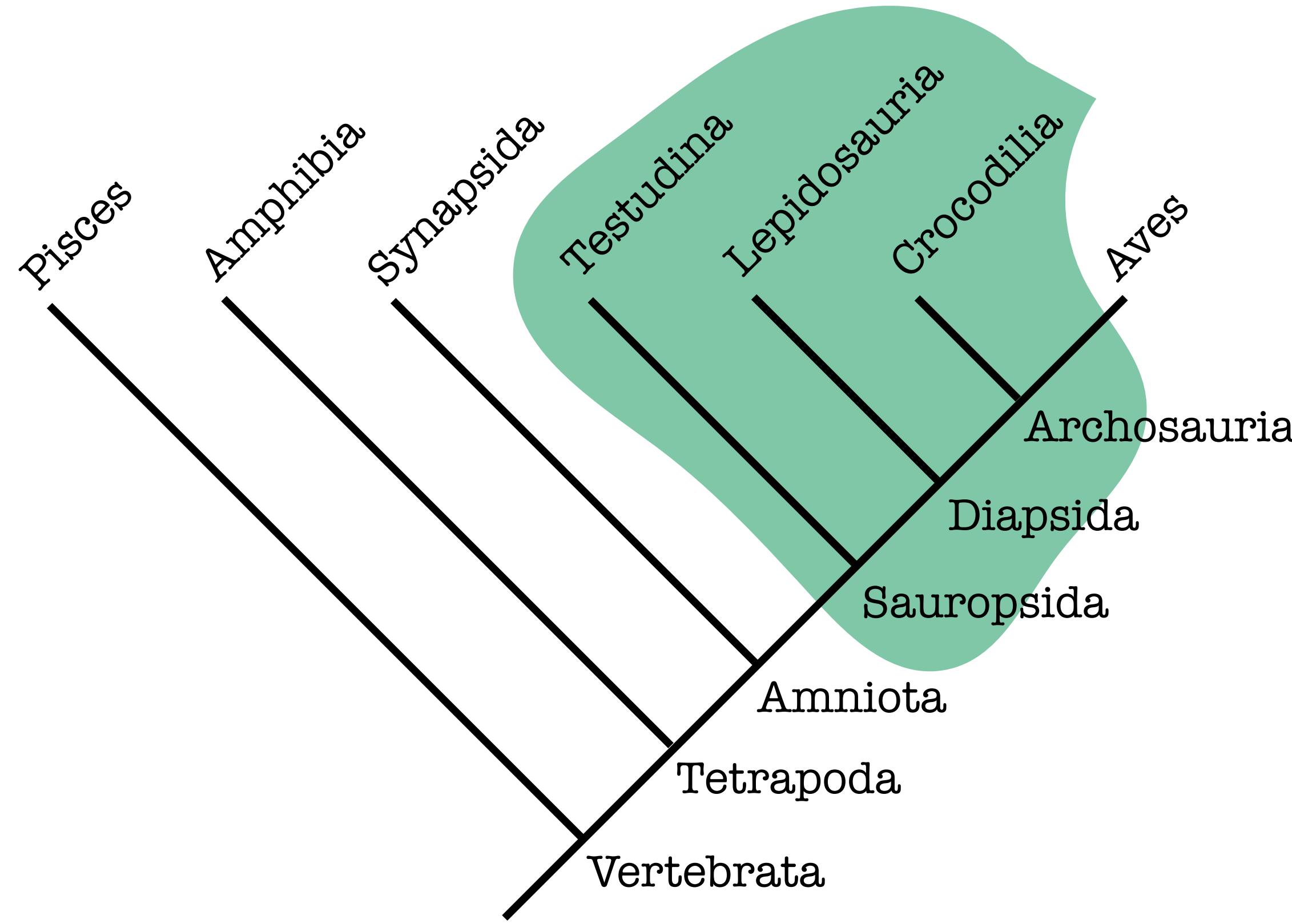
Parcialmente resuelto

Politomía

Árboles: terminología

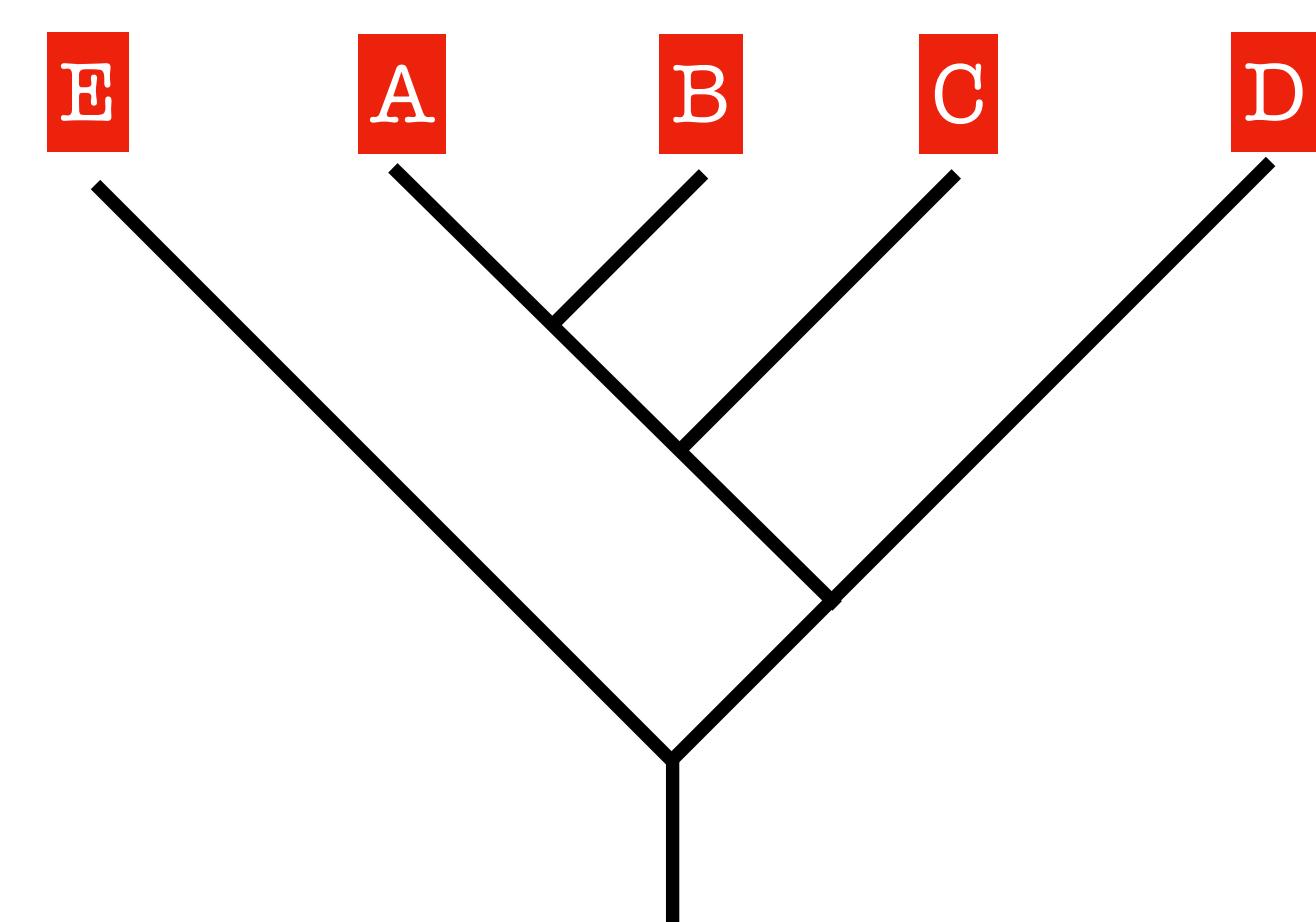
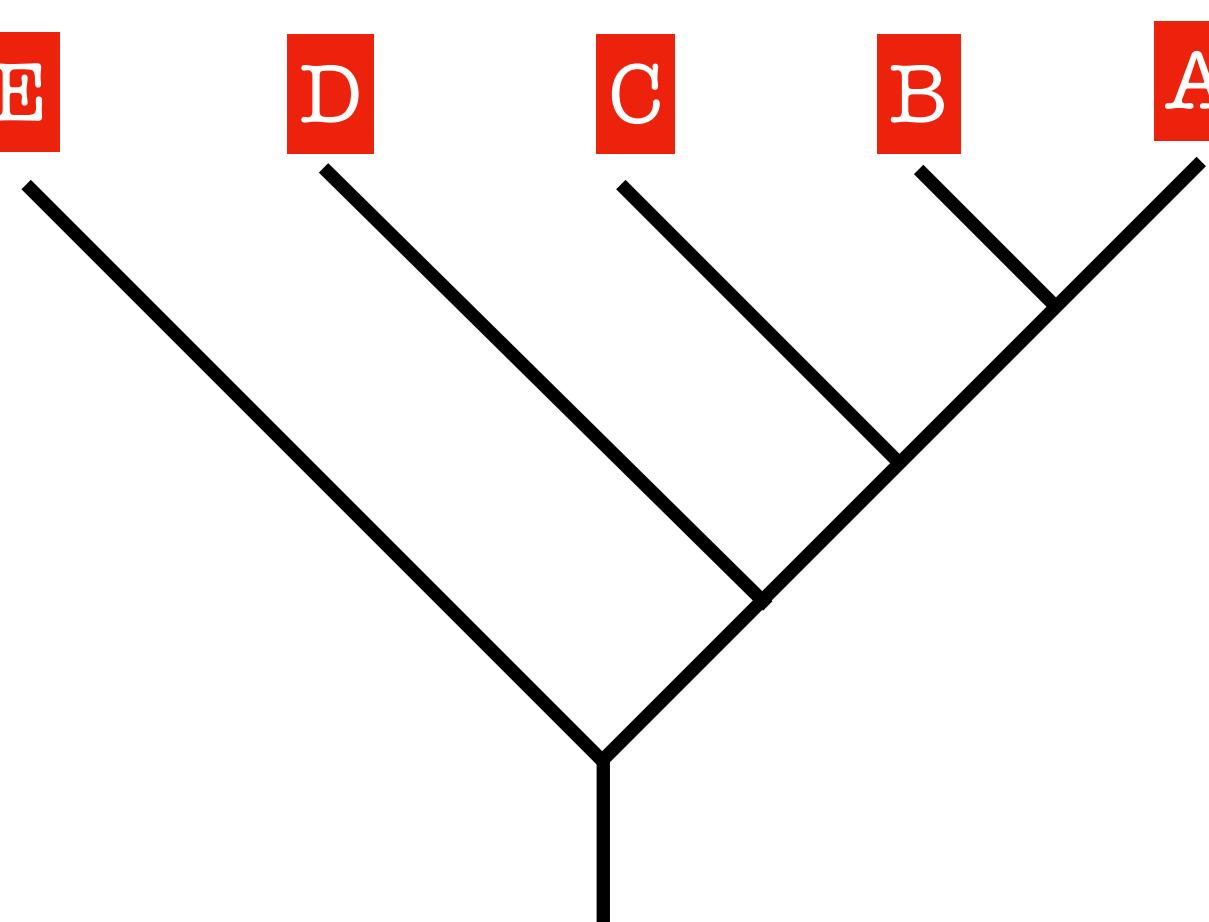
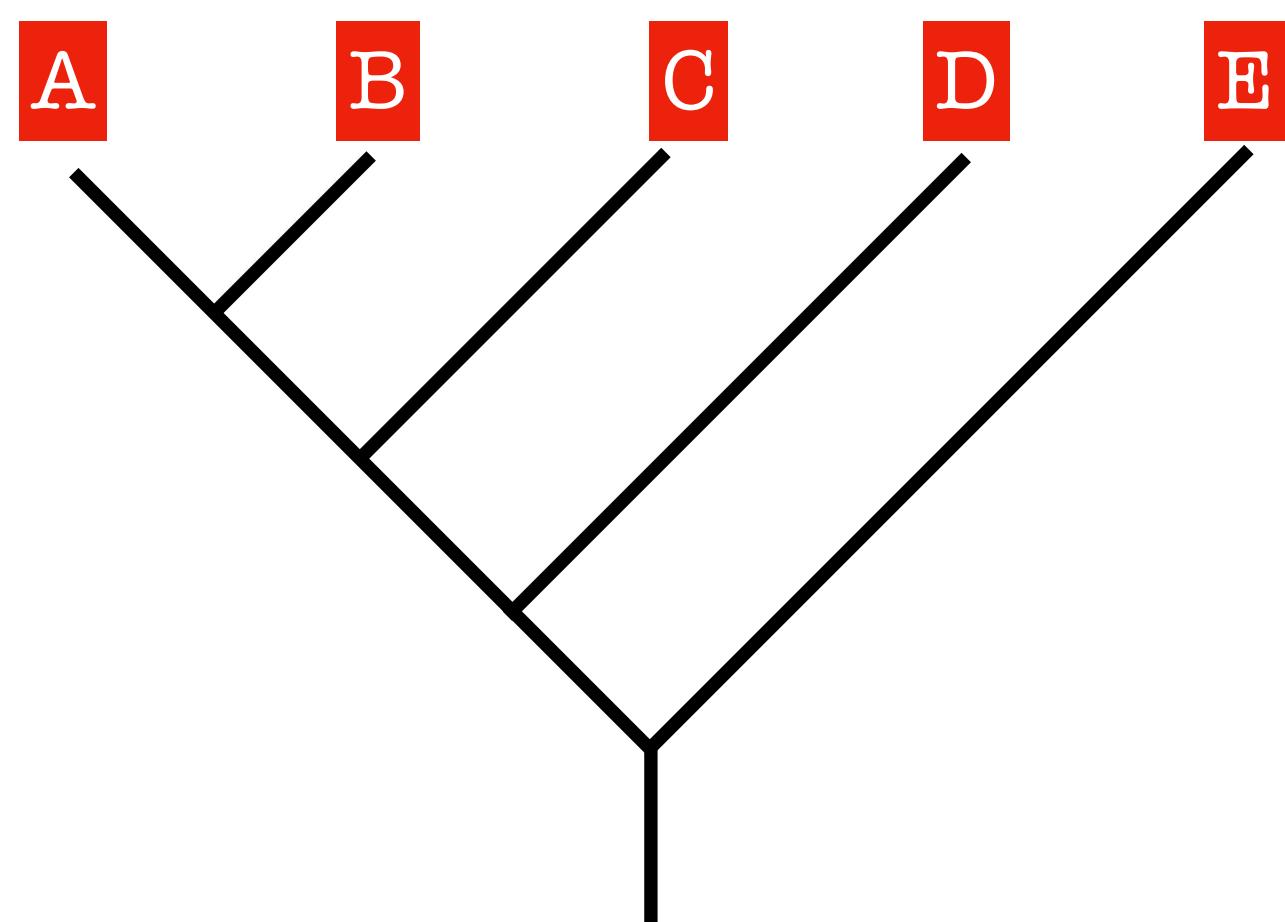


Árboles: terminología

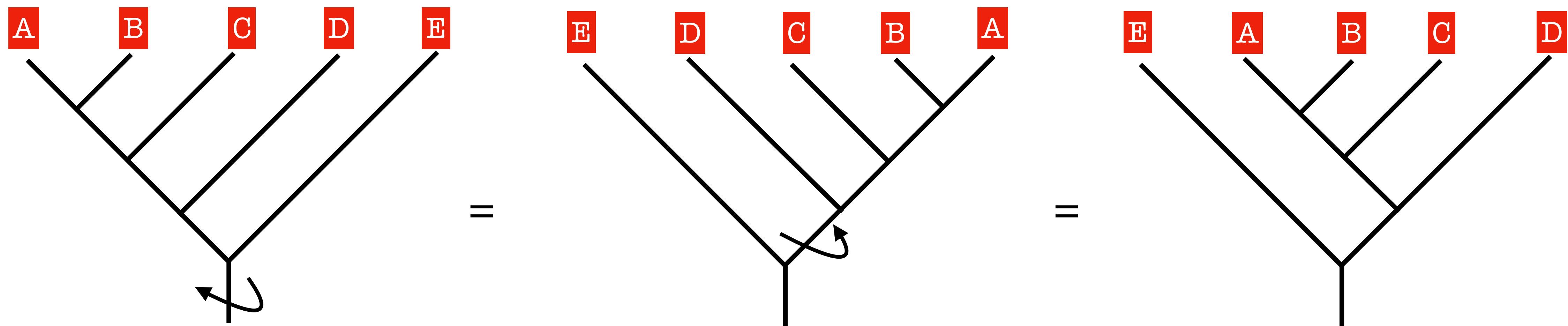


Los **reptiles** no son un grupo monofilético ... a menos que incluya a las aves!

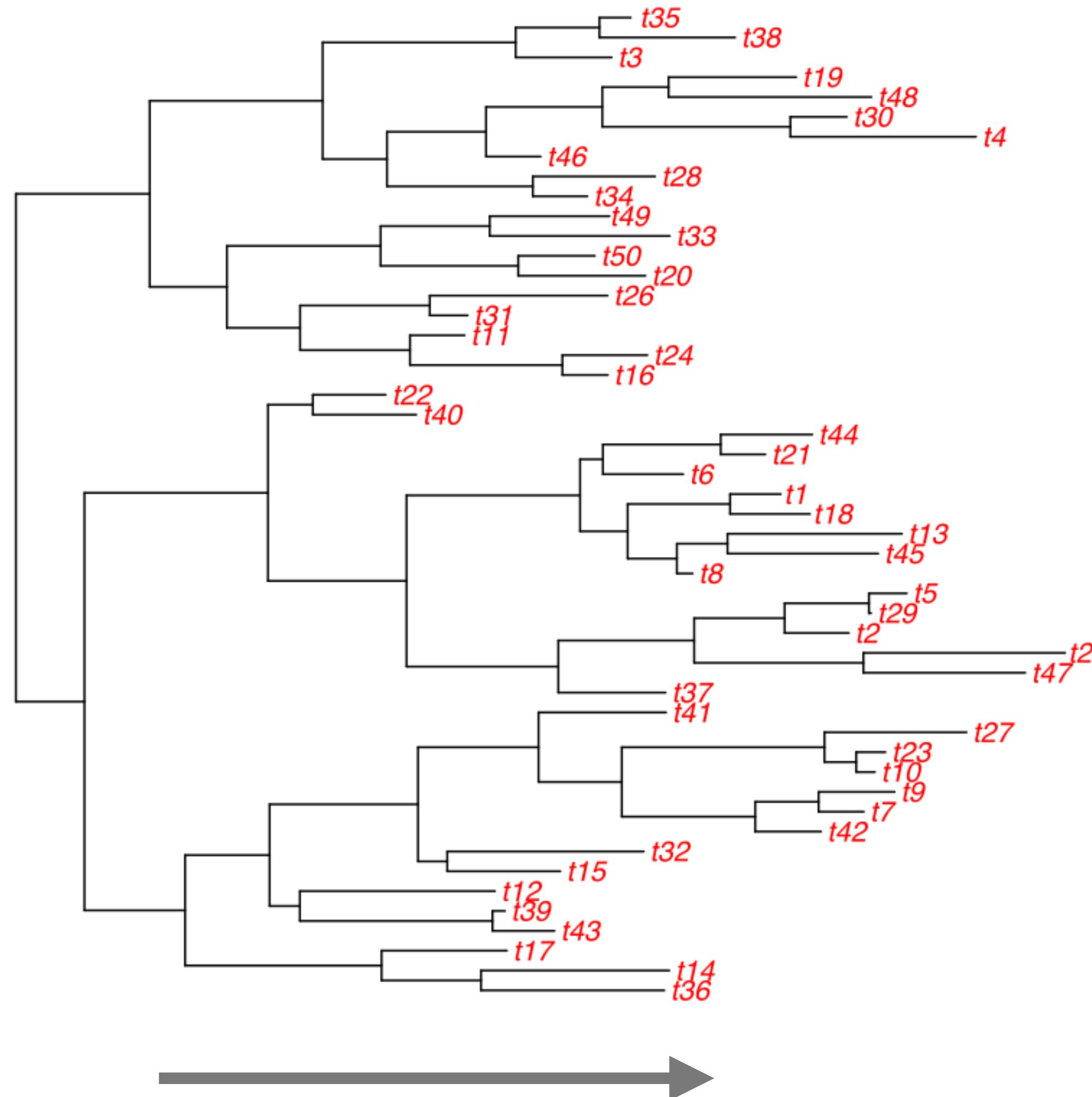
Árboles: representación



Árboles: representación

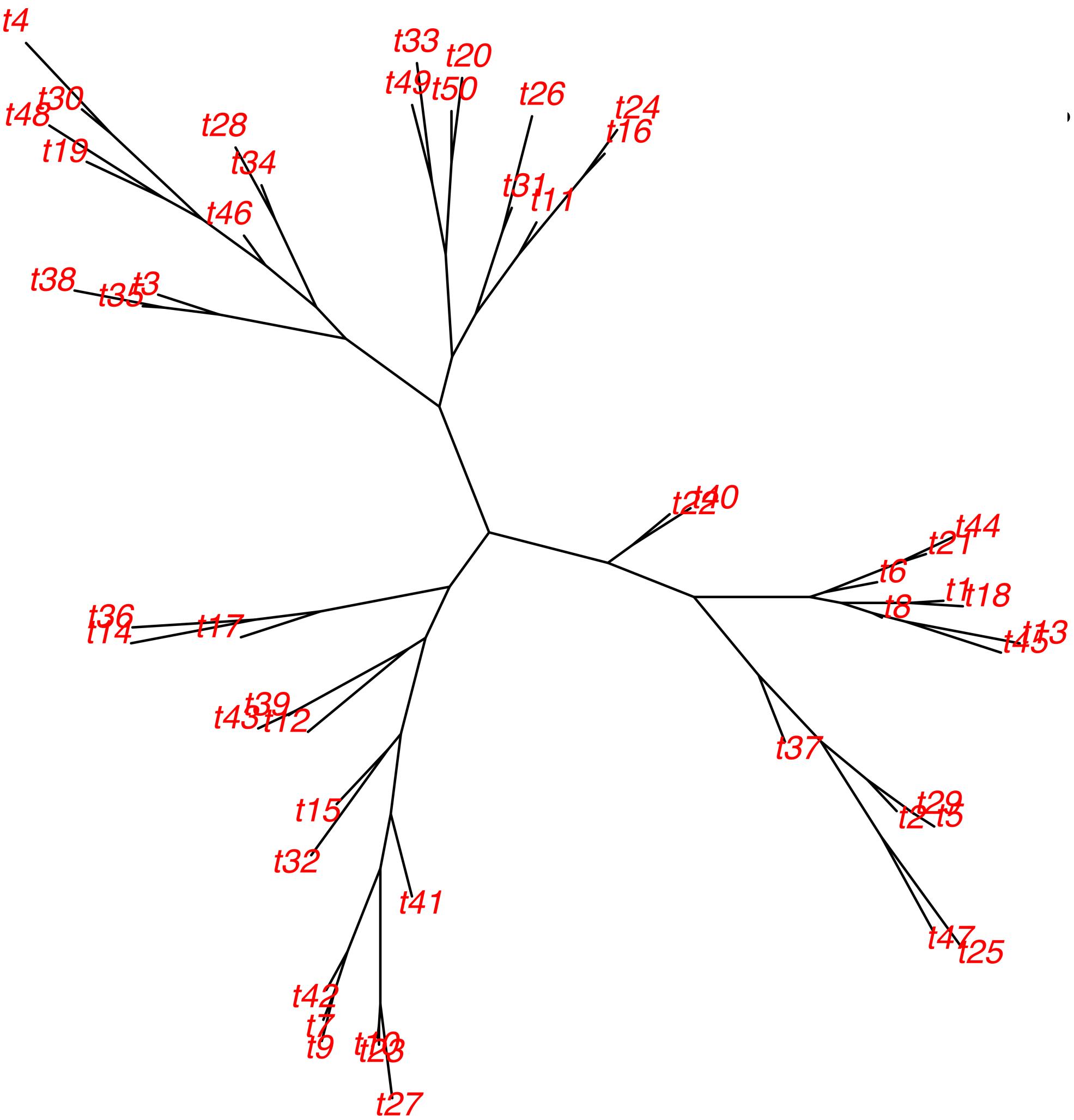


Árboles: con raíz vs. sin raíz



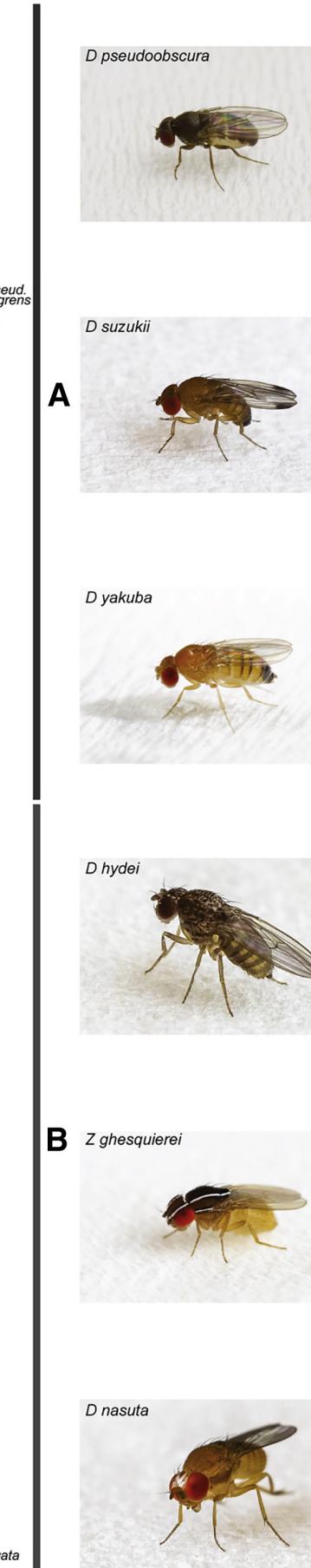
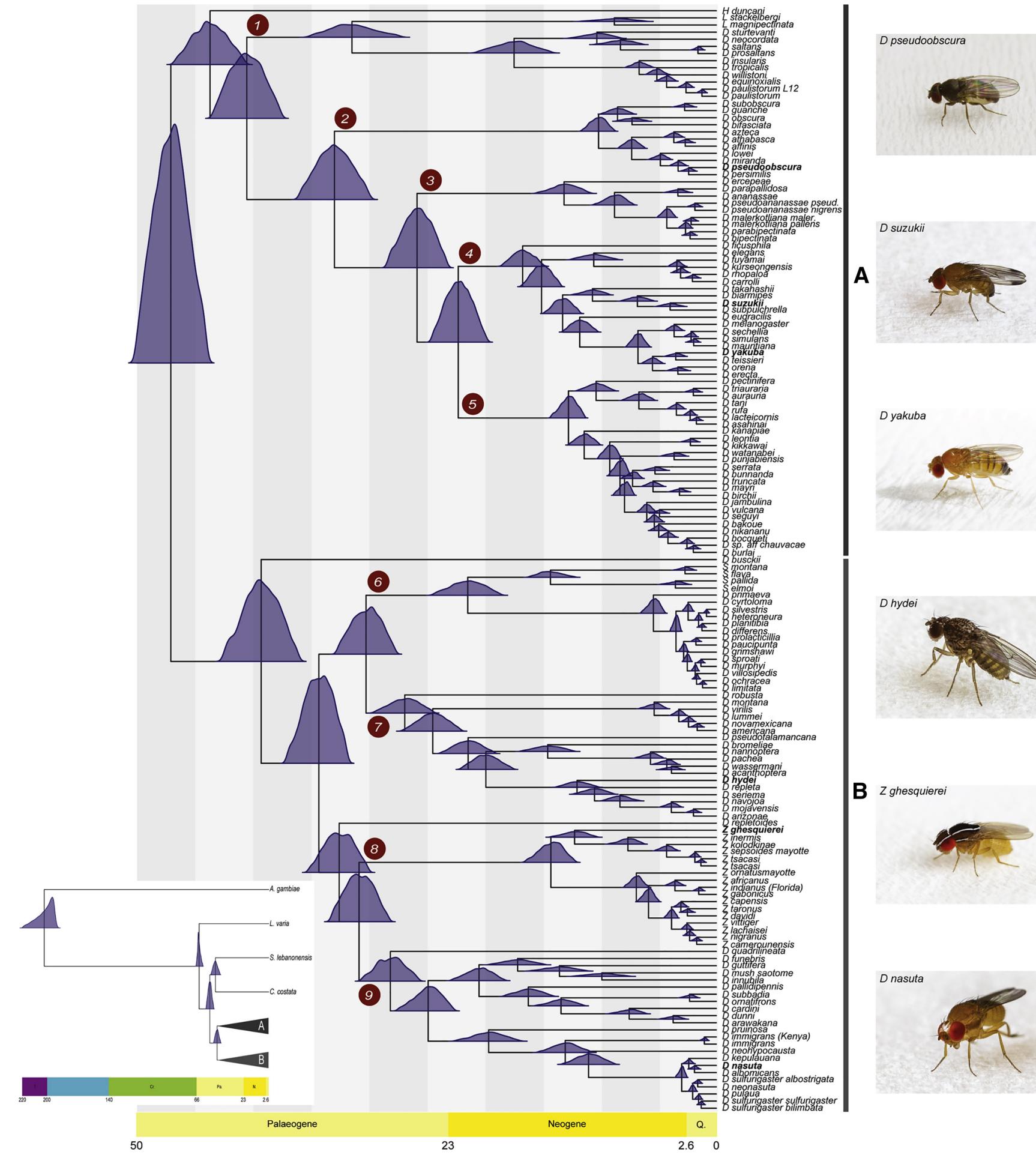
- Los árboles con raíz tienen un nodo raíz específico, que representa al ancestro común de todos los organismos del árbol
- La raíz representa un punto en el tiempo anterior a cualquier otro nodo del árbol
- Un árbol enraizado tiene dirección (los nodos se pueden ordenar por “anteriores” o “posteriores”)
- En un árbol enraizado, la distancia entre dos nodos se representa únicamente a lo largo del eje temporal (el segundo eje solo ayuda a distribuir las terminales)

Árboles: con raíz vs. sin raíz



- Los árboles sin raíz no tienen un nodo raíz específico y solo muestran el patrón de ramificación de las relaciones evolutivas entre taxones, sin ninguna información sobre su ancestro común
- La distancia a lo largo de las ramas directamente representa las distancias entre nodos.

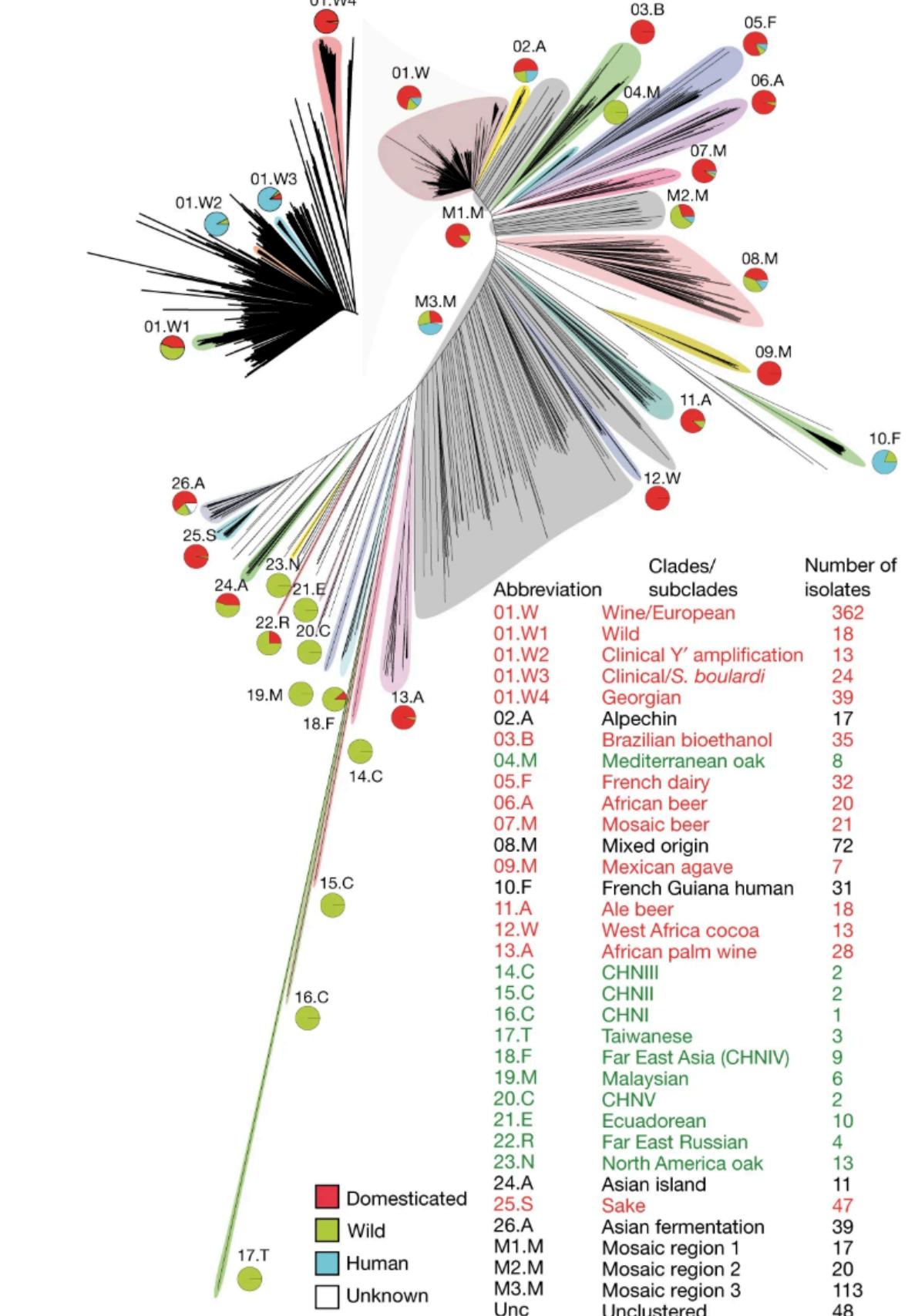
Tipos de árboles filogenético



B

Fig. 1: Neighbour-joining tree built using the biallelic SNPs.

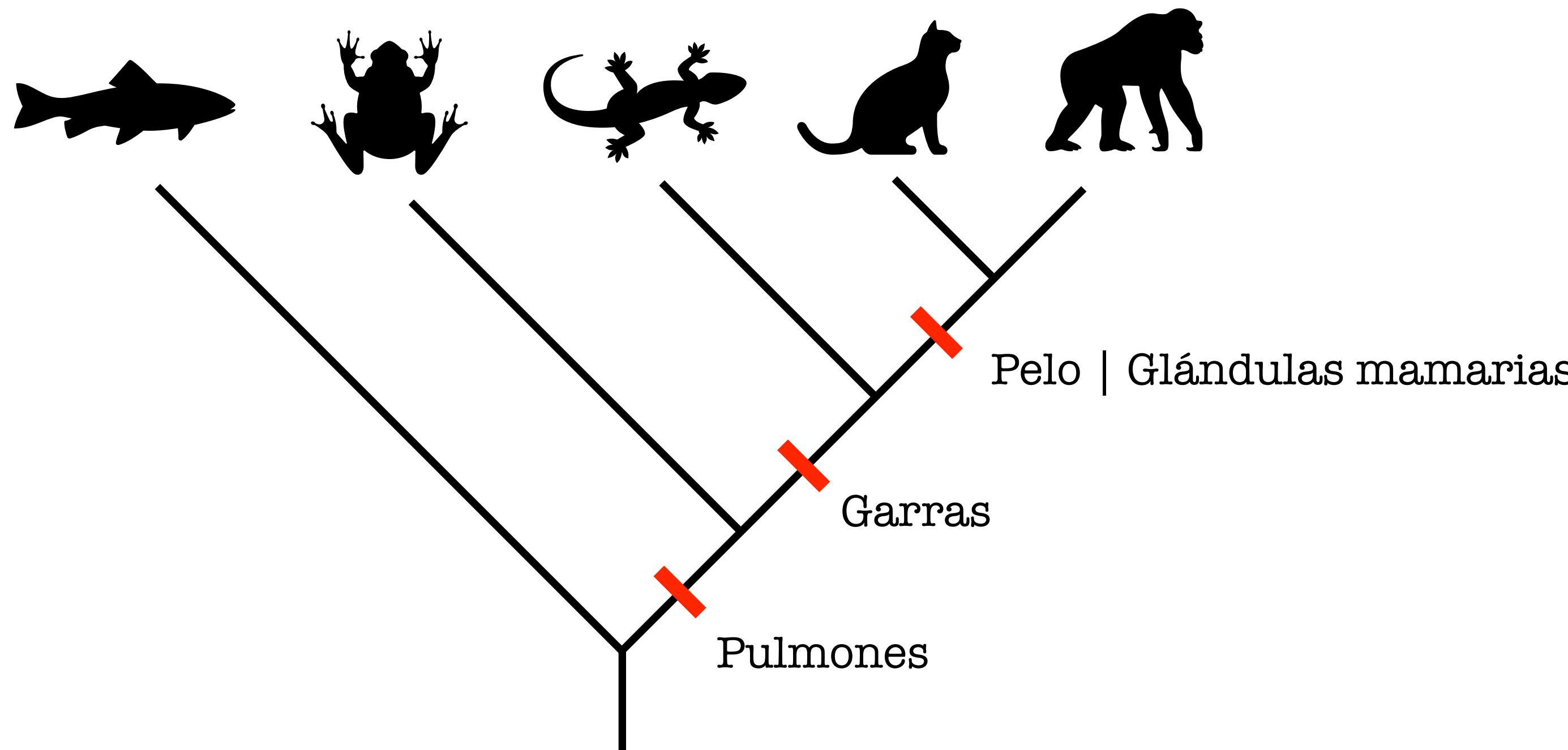
From: [Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates](#)



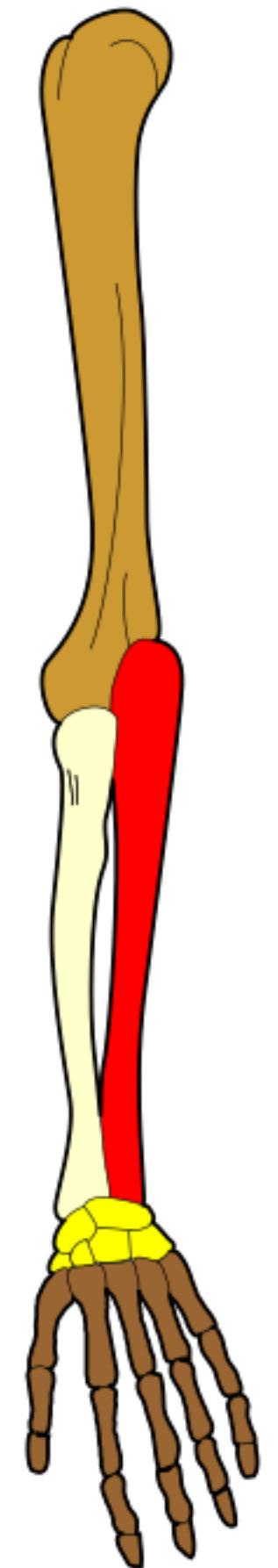
We identified 26 clades (numbered clockwise from 1 to 26) and three mosaic groups (M1–M3). The pie charts represent the ecological origins of the clade: domesticated (red), wild (green) and human (cyan). The colour of the clade name indicates its assignment: domesticated (red) and wild (green). The top left inset represents a magnification of the wine/European clade with four major subclades highlighted.

¿Cómo reconstruimos un árbol filogenético a partir de datos?

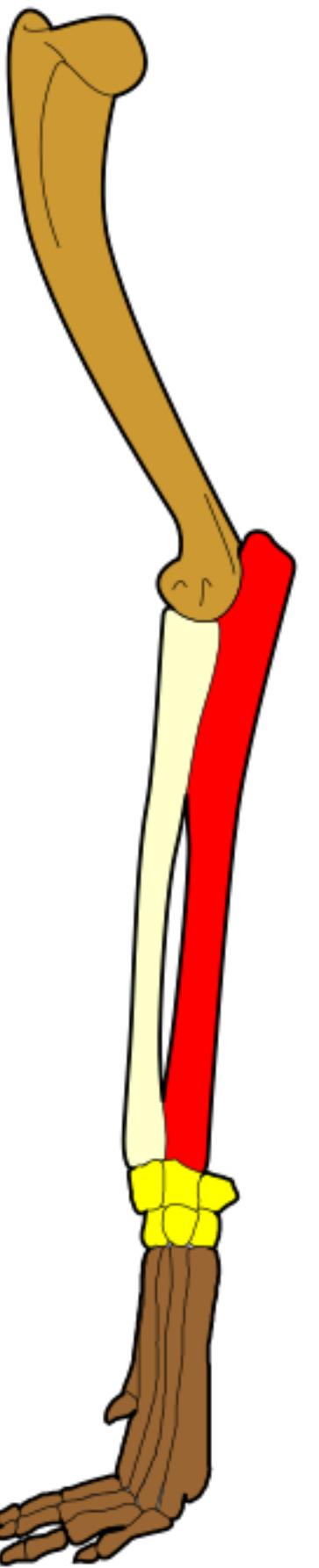
Reconstrucción filogenética: datos obtenidos en la actualidad



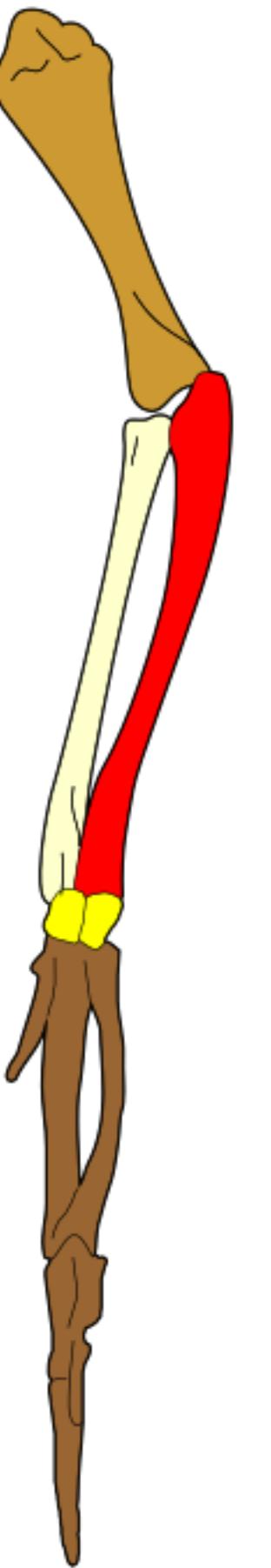
Homología vs. Homoplasia



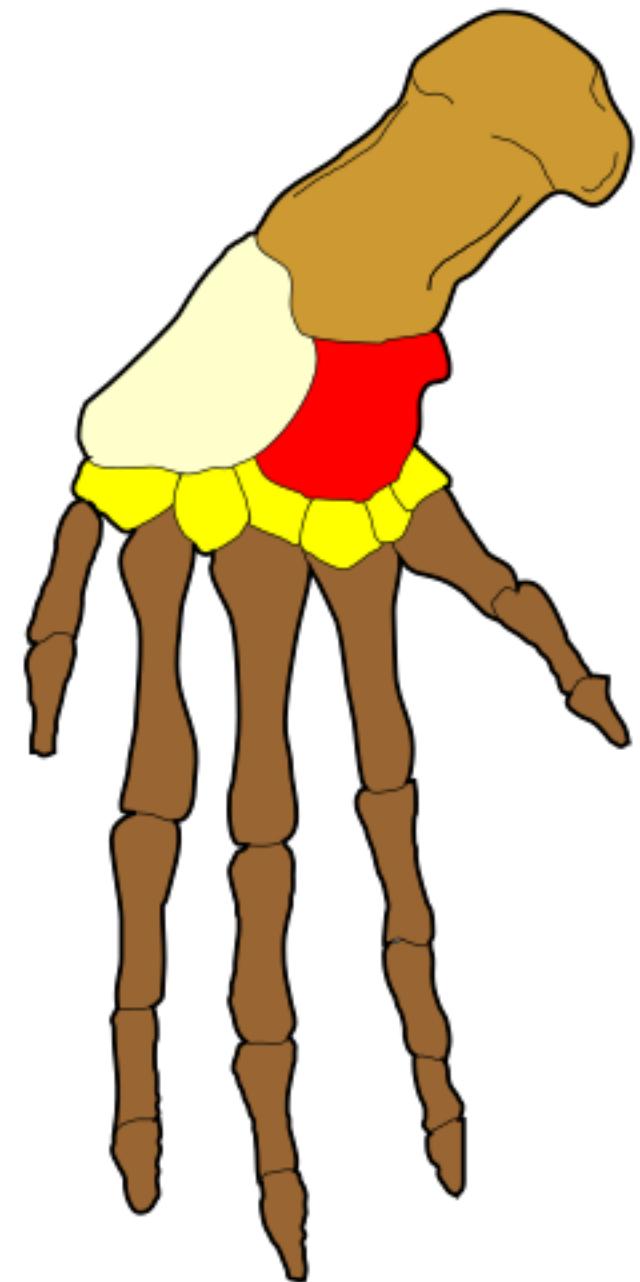
Humano



Perro



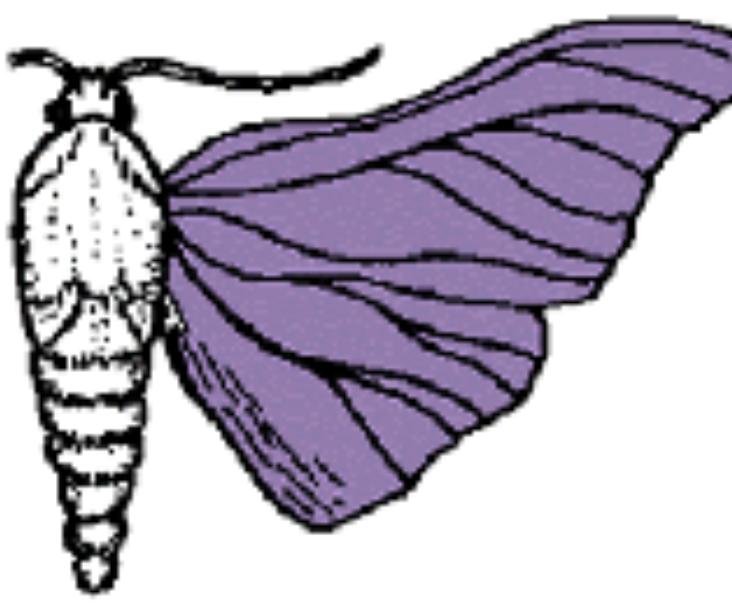
Pájaro



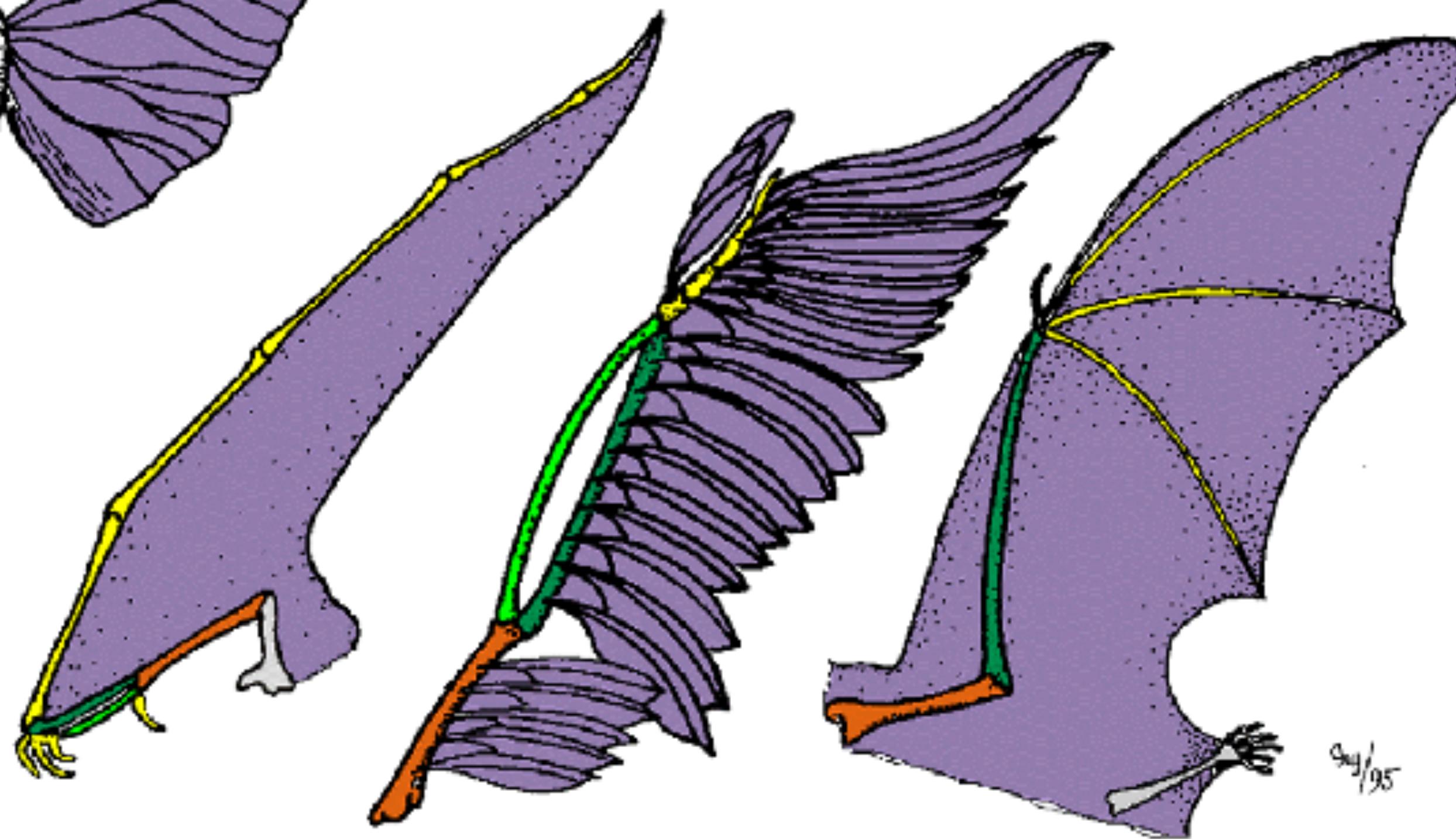
Ballena

Homología vs. Homoplasia

Insecto



Pterodactilo



Ave

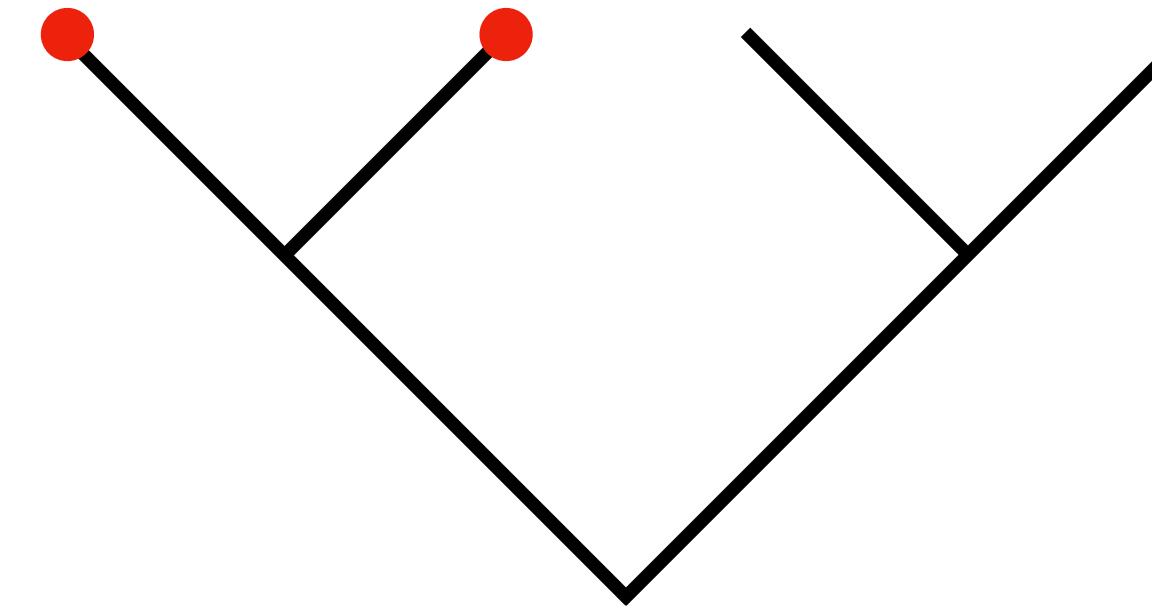


Murciélagos

94/95

Homología vs. Homoplásia

Los caracteres homólogos son los que dan información para reconstruir la filogenia



Homología: caracteres similares heredados de un ancestro común

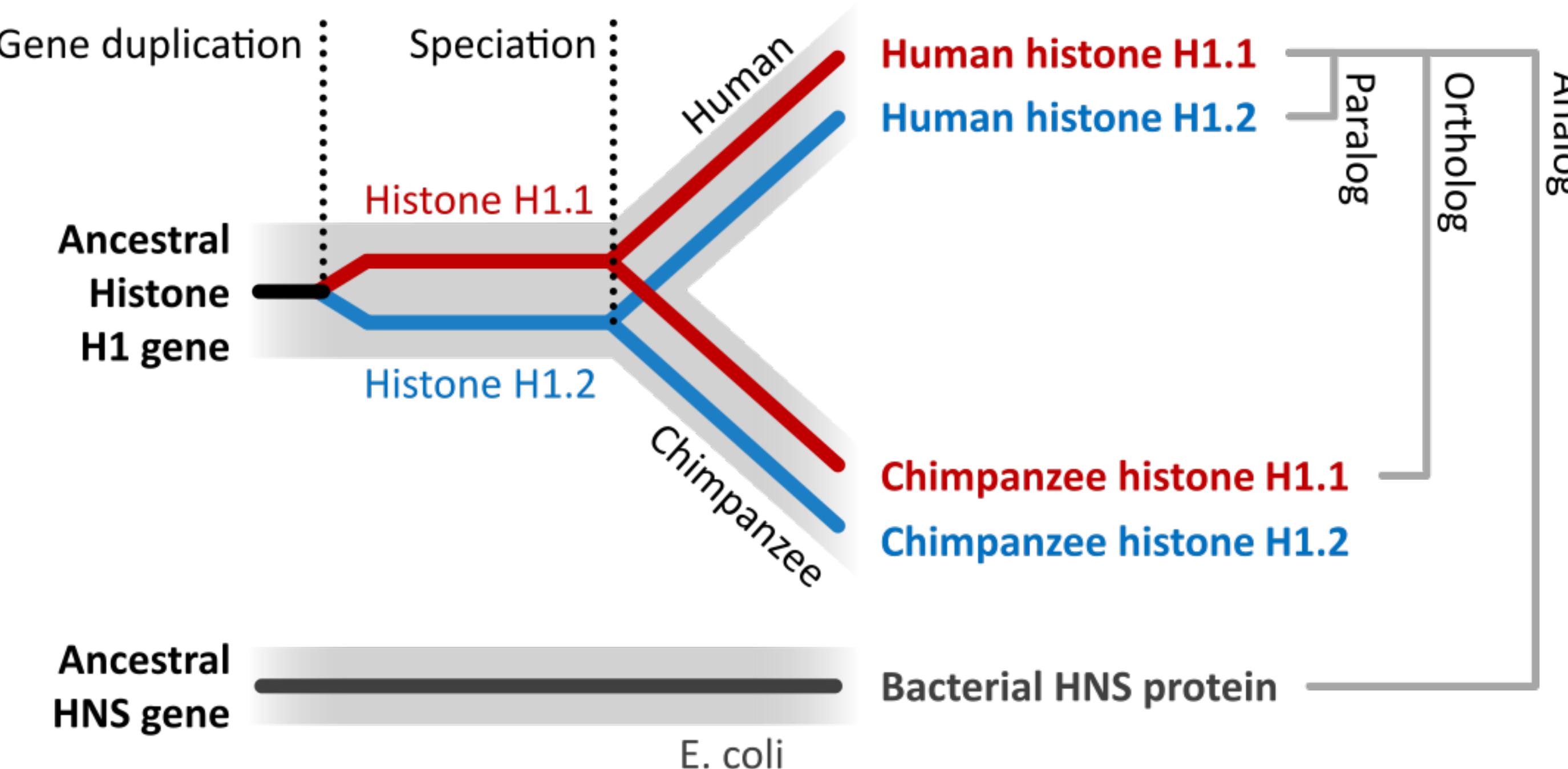
Homoplásia: caracteres similares no son directamente originados de un ancestro común (evolución convergente)

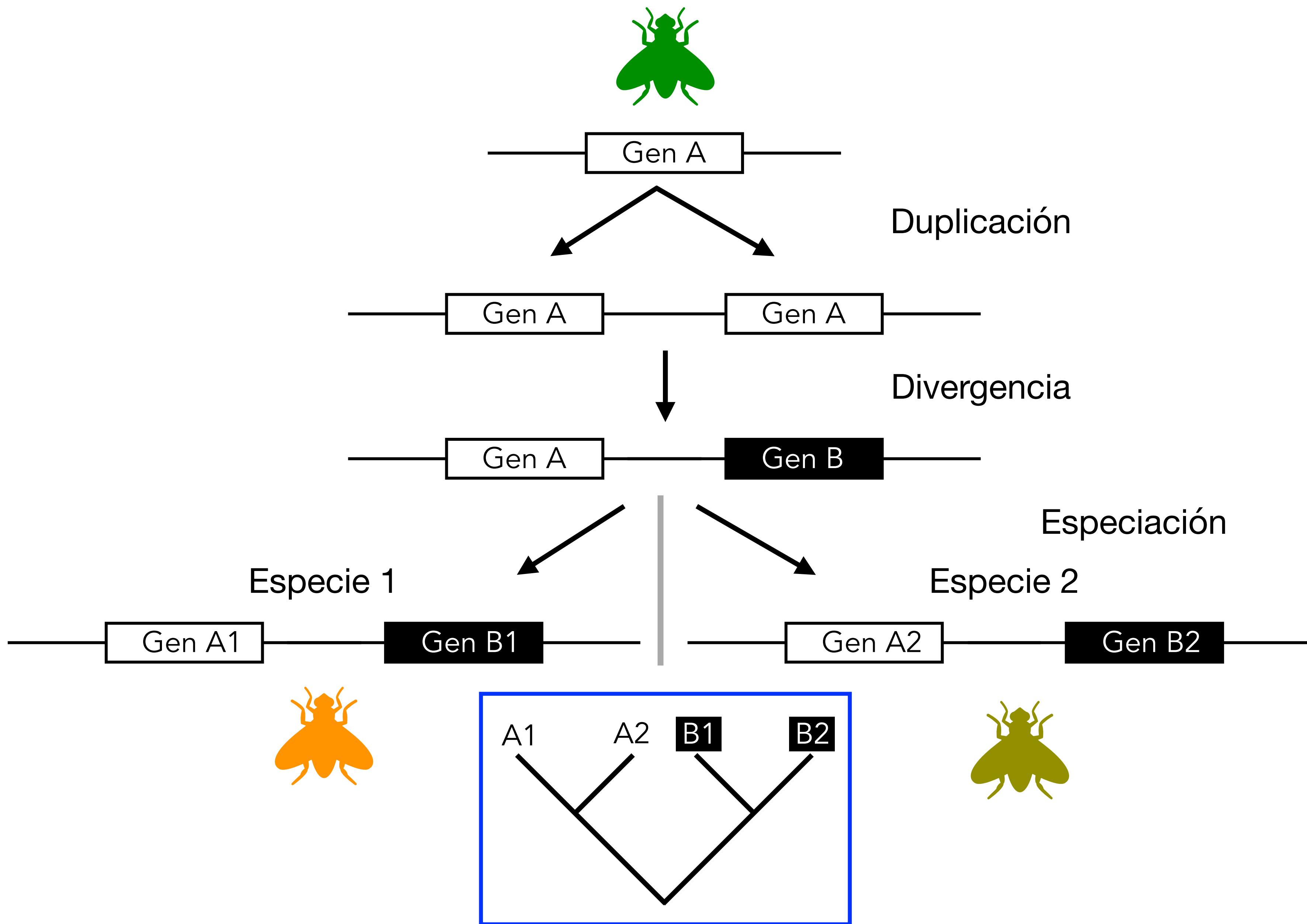
A nivel molecular ...

- **Homología:** dos o más secuencias de proteínas o ácidos nucleicos (ADN/ARN) son similares entre sí debido a que presentan un mismo origen evolutivo
- Normalmente concluimos que dos secuencias son homólogas por el alto grado de similitud que presentan

A nivel molecular ...

- **Homología:** dos o más secuencias de proteínas o ácidos nucleicos (ADN/ARN) son similares entre sí debido a que presentan un mismo origen evolutivo
- Normalmente concluimos que dos secuencias son homólogas por el alto grado de similitud que presentan





Filogenia molecular

A	A G C G T T G G G C A A
B	A G C G T T T G G C A A
C	A G C T T T G T G C A A
D	A G C T T T T T G C A A

1 2 3

- Información obtenida a partir de secuencias de ADN o proteínas
- Los caracteres homólogos son inferidos a partir de un alineamiento
- Las distintas bases en una posición son los estados de ese carácter
- Otros datos moleculares, como ser: ausencia o presencia de sitios de restricción, reactividad cruzada de anticuerpos, también pueden ser considerados (no son muy utilizados en la actualidad)

Pasos para conseguir un árbol filogenético a partir de datos moleculares

1) Elección del marcador molecular

- identificación de ortólogos
- variabilidad (aa vs. nt)

2) Alineamiento

3) Curado del alineamiento (automático // manual) & evaluación del alineamiento

4) Selección del modelo evolutivo

5) Análisis filogenético (agrupamiento y estimación de las distancias)

- Máxima parsimonia
- **Métodos por distancias**
- **Máxima Verosimilitud**
- Bayesianos

Alineamientos

Taxon	Unaligned	Aligned	Trimmed
1	ACTGCGTTAGGTCTAGCC	-----ACT-GCGTTAGG-TCTAGCC--	-----ACT-GCGTTAGG-TCTAGCC--
2	GATCTACTGCTTTAGGTTGAGCC	→ GATCTACT-GCTTTAGG-TTGAGCC--	→ GATCTACT-GCTTTAGG-TTGAGCC--
3	ACTGCTCTAGCACTGAGCCCCA	-----ACT-GCTCTAGCACTGAGCCCCA	-----ACT-GCTCTAGCACTGAGCCCCA
4	ACTTGGCGTAGCCGGAGGCC	-----ACTTGGCGTAGC-CGGAGGCC-	-----ACTTGGCGTAGC-CGGAGGCC-

Métodos Basados en Matrices de Distancia

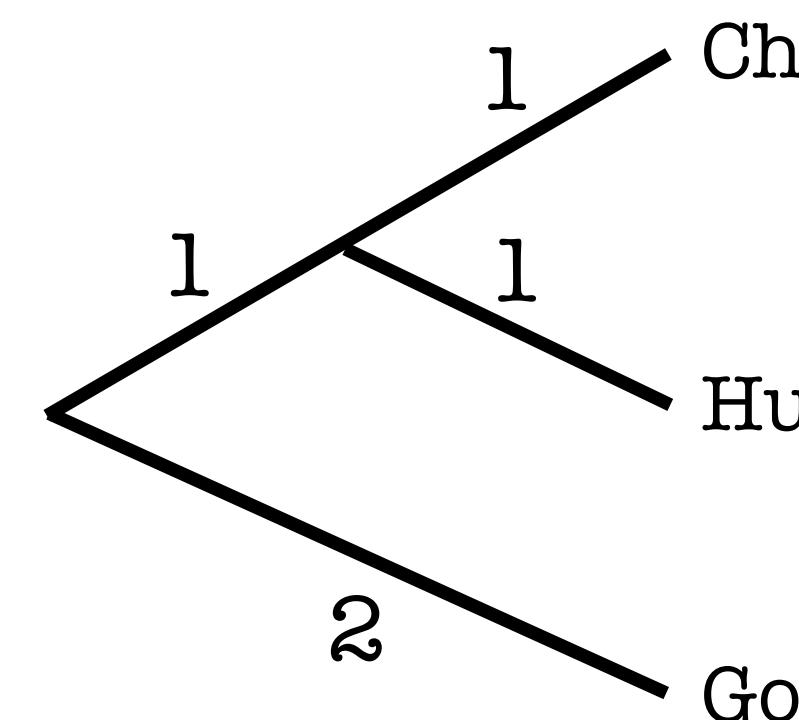
Gorila : ACGT**CGTA**
Humano : ACGTTCCCT
Chimpance : ACGTT**T**C**G**

↓↓↓↓
↑↑

1) Construcción de alineamiento

	Go	Hu	Ch
Go	-	4	4
Hu		-	2
Ch			-

2) Construimos una tabla con todas las **diferencias pareadas** entre las secuencias (matriz de distancia).

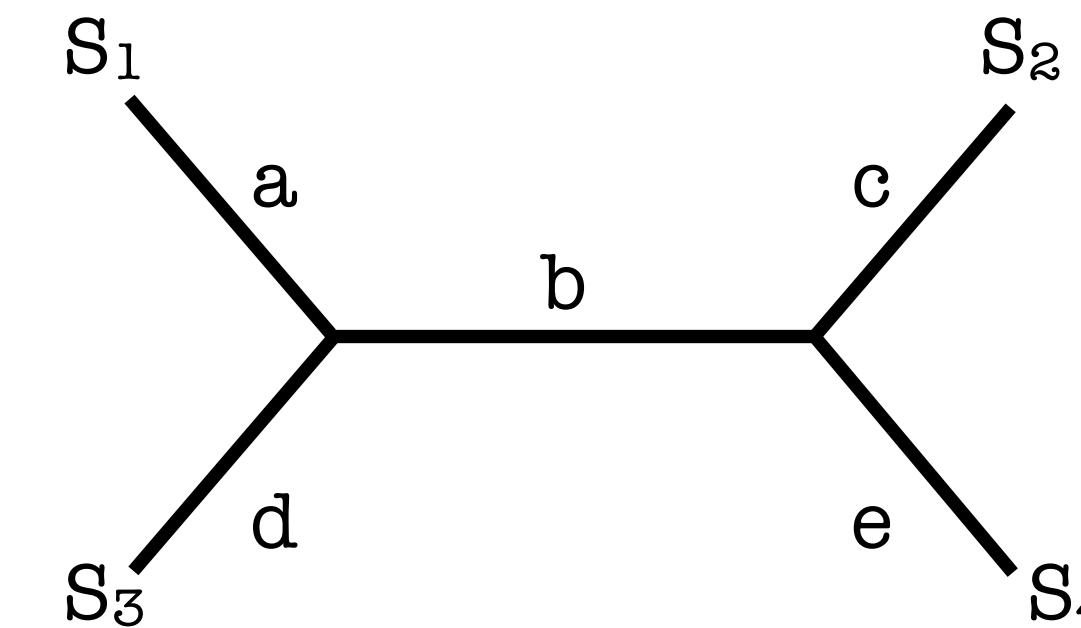


3) Construimos un árbol a partir de la matriz de distancias de forma que coincidan con la matriz.

Métodos por distancia: búsqueda del árbol óptimo.

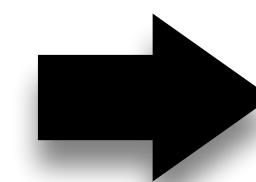
	S ₁	S ₂	S ₃	S ₄
S ₁	-	D ₁₂	D ₁₃	D ₁₄
S ₂		-	D ₂₃	D ₂₄
S ₃			-	D ₃₄
S ₄				-

Distancias Observadas



Distancias patrísticas (o cofenéticas)

Objetivo: Encontrar un árbol tal que las distancias cofenéticas sean lo más parecidas posibles a las distancias observadas

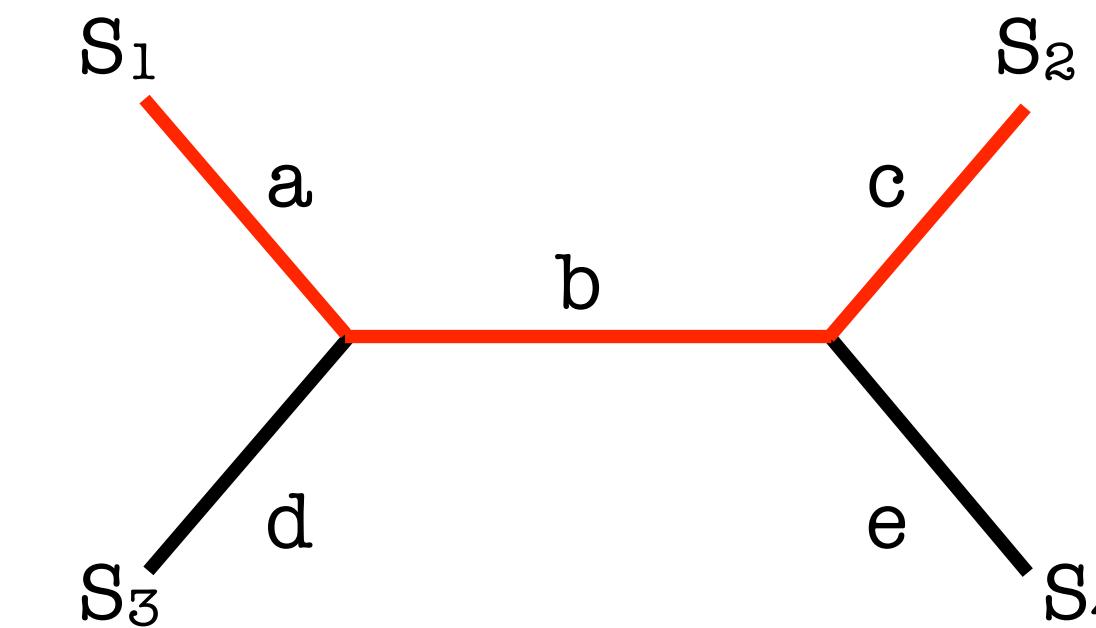


$$\begin{aligned}D_{12} &\approx d_{12} = a + b + c \\D_{13} &\approx d_{13} = a + d \\D_{14} &\approx d_{14} = a + b + e \\D_{23} &\approx d_{23} = d + b + c \\D_{24} &\approx d_{24} = c + e \\D_{34} &\approx d_{34} = d + b + e\end{aligned}$$

Métodos por distancia: búsqueda del árbol óptimo.

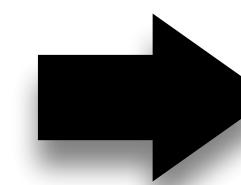
	S ₁	S ₂	S ₃	S ₄
S ₁	-	D ₁₂	D ₁₃	D ₁₄
S ₂		-	D ₂₃	D ₂₄
S ₃			-	D ₃₄
S ₄				-

Distancias Observadas



Distancias patrísticas (o cofenéticas)

Objetivo: Encontrar un árbol tal que las distancias cofenéticas sean lo más parecidas posibles a las distancias observadas



$$D_{12} \approx d_{12} = a + b + c$$

$$D_{13} \approx d_{13} = a + d$$

$$D_{14} \approx d_{14} = a + b + e$$

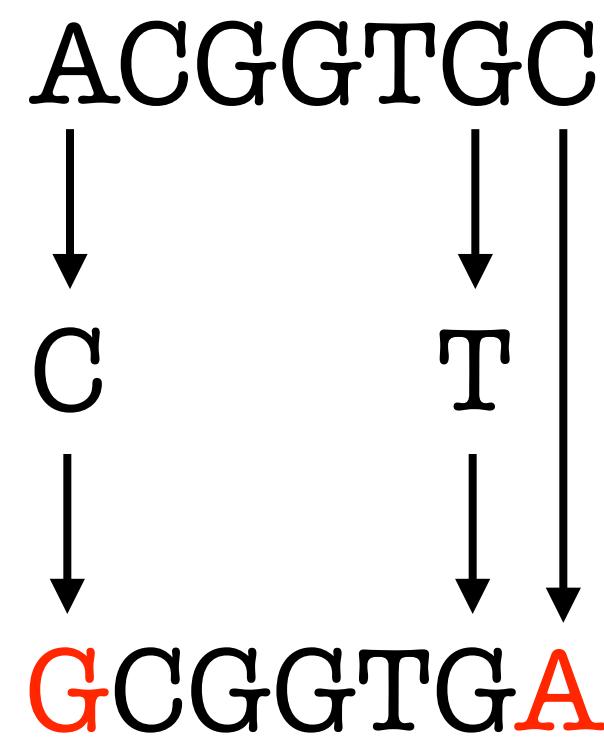
$$D_{23} \approx d_{23} = d + b + c$$

$$D_{24} \approx d_{24} = c + e$$

$$D_{34} \approx d_{34} = d + b + e$$

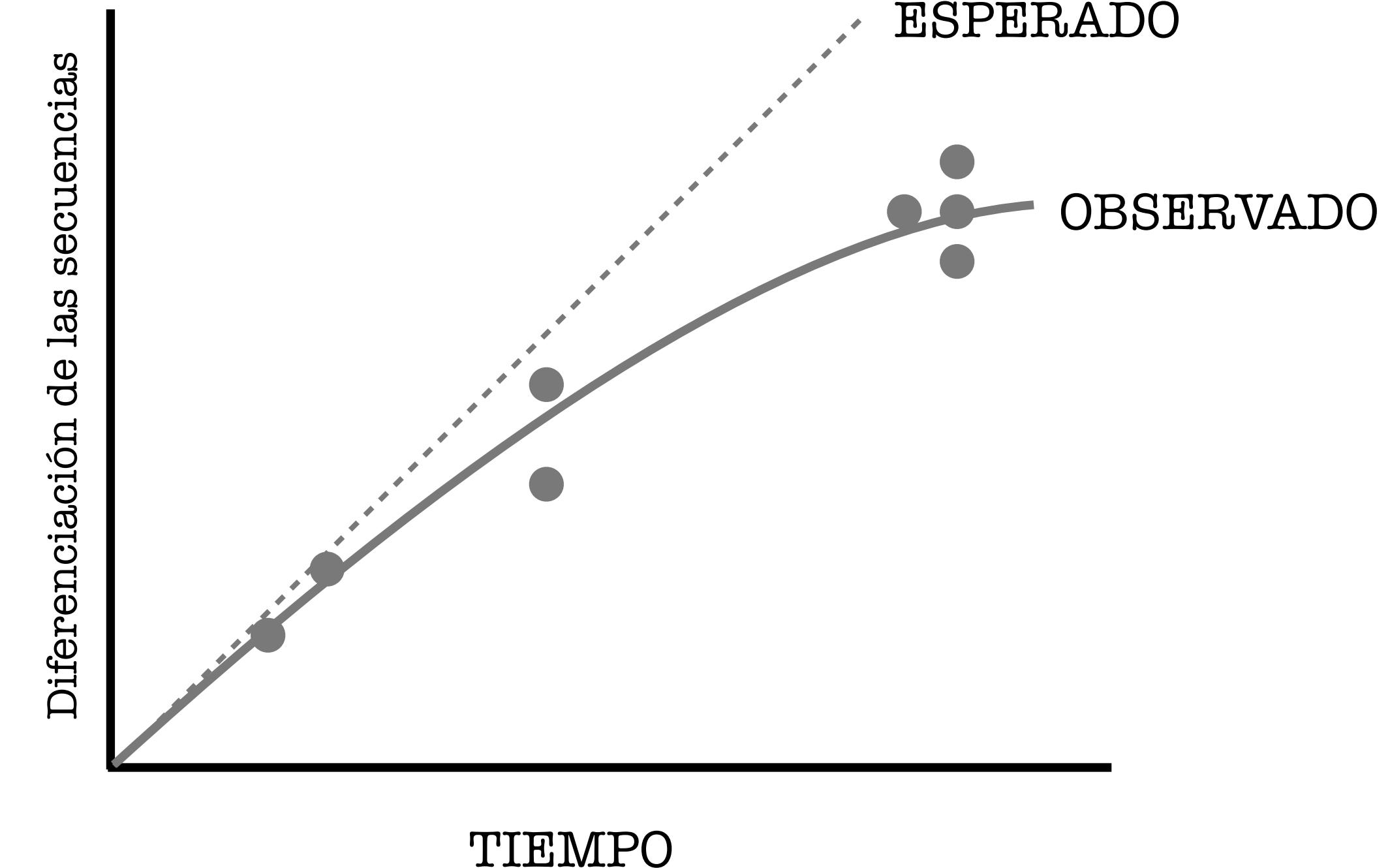
Las distancias observadas no siempre igualan las estimadas...

Mutaciones superpuestas



- Cambios que ocurrieron en la evolución: 5
- Cambios observados: 2

La distancia (casi) siempre se subestima



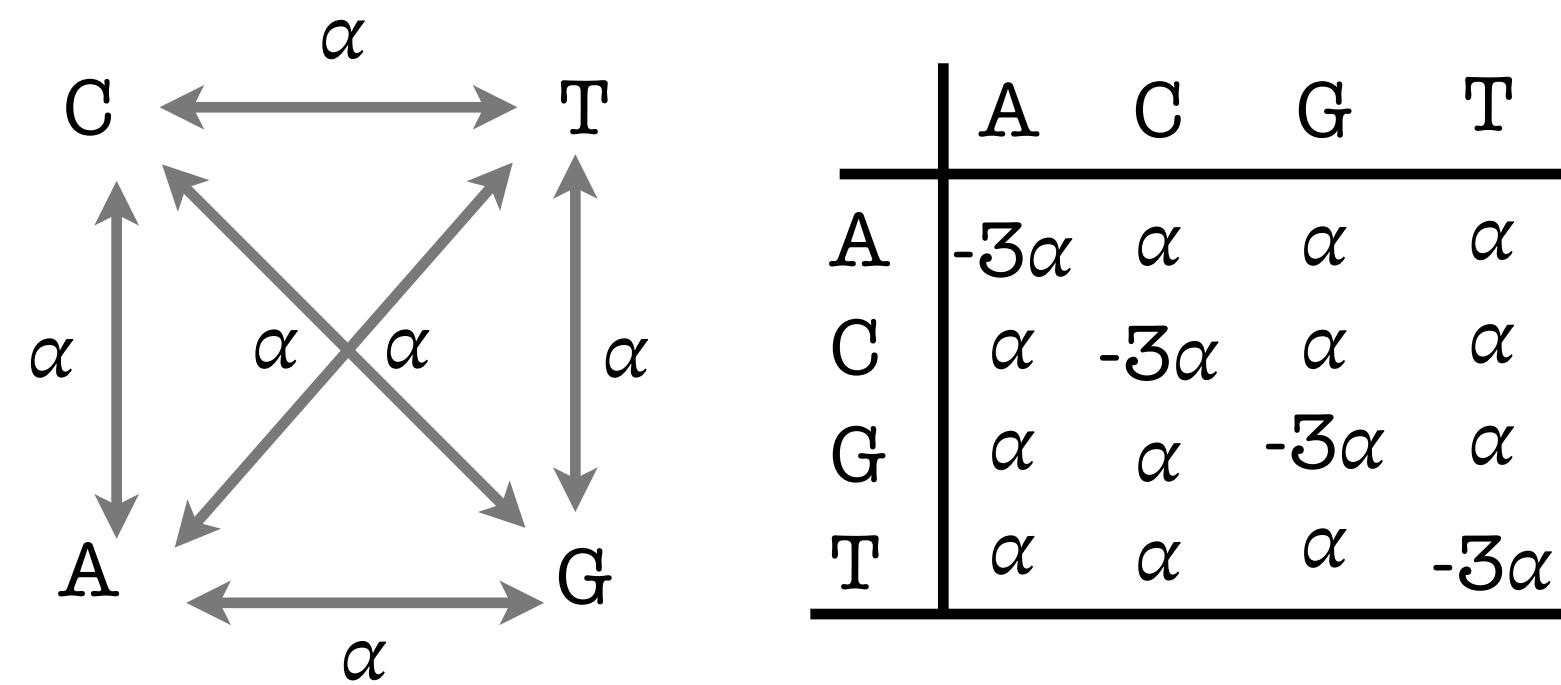
Corrección basada en modelos para las mutaciones superpuestas

Objetivo

- Inferir el número real de eventos evolutivos (la distancia real) basándose en:
 - ▶ los datos observados (alineamiento de secuencias)
 - ▶ un modelo de cómo se produce la evolución.

Corrección basada en modelos para las mutaciones superpuestas

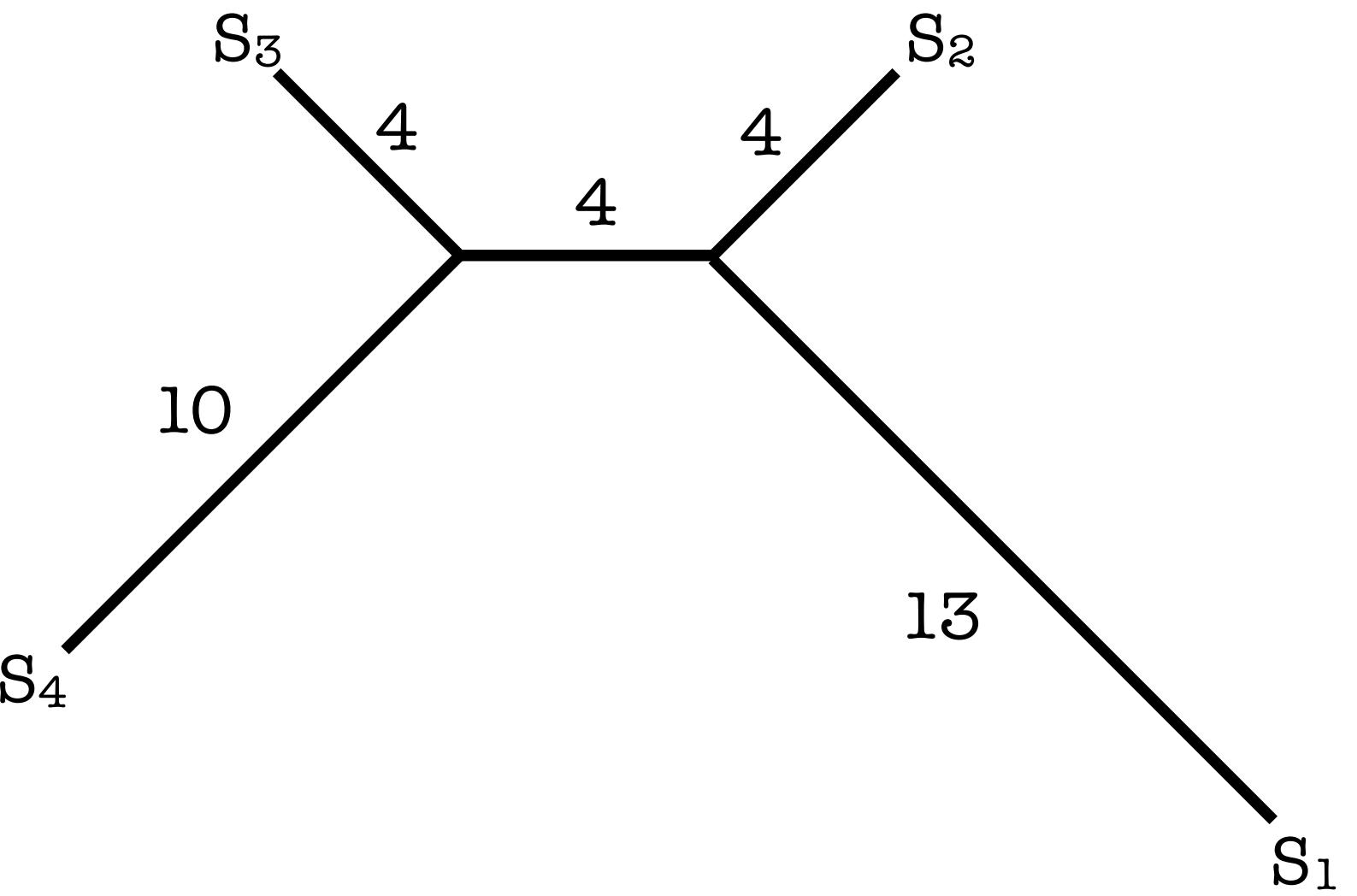
Modelo de Jukes & Cantor



- Asume que los cuatro nucleótidos tienen la misma frecuencia ($f = 0,25$).
- Asume que las 12 tasas de sustitución son iguales.
- En este modelo, la distancia corregida es: $D_{JC} = -\frac{3}{4} \ln(1 - \frac{4}{3} D_{OBS})$
- Por ejemplo: $D_{OBS} = 0,42 \rightarrow D_{JC} = 0,62$

Algoritmo Neighbor Joining

D_{ij}	S ₁	S ₂	S ₃	S ₄
S ₁	-	17	21	27
S ₂		-	12	18
S ₃			-	14
S ₄				-



Maxima Verosimilitud

Enfoque I

- **Punto de partida**

- Datos observados y un modelo probabilístico sobre cómo se produjeron esos datos.
- Tener un modelo probabilístico de un proceso significa que podemos calcular la probabilidad de cualquier resultado posible (dado un conjunto de valores específicos para los parámetros del modelo).

- **Ejemplo**

- **Datos:** resultado de lanzar una moneda 10 veces: **7 caras, 3 cruces**.
- **Modelo:** la moneda tiene una probabilidad p de que salga cara y $1-p$ de que salga cruz.
- La probabilidad de observar c caras en n lanzamientos es:

$$\mathbb{P}_{(c \text{ caras})} = \binom{n}{c} p^c (1 - p)^{n-c}$$

- **Objetivo**

- Encontrar la mejor estimación de los valores de los parámetros (desconocidos) basándose en las observaciones
- Aquí el único parámetro es p

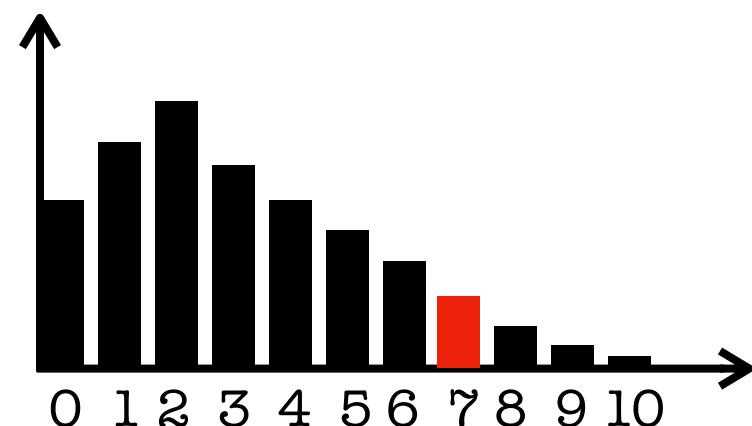
Maxima Verosimilitud

Por lo tanto ... los
parámetros del
modelo!

Enfoque II

- Verosimilitud (Modelo) = Probabilidad (Datos | Modelo)
- Máxima verosimilitud: La mejor estimación es el conjunto de valores de los parámetros que ofrece la mayor verosimilitud posible.

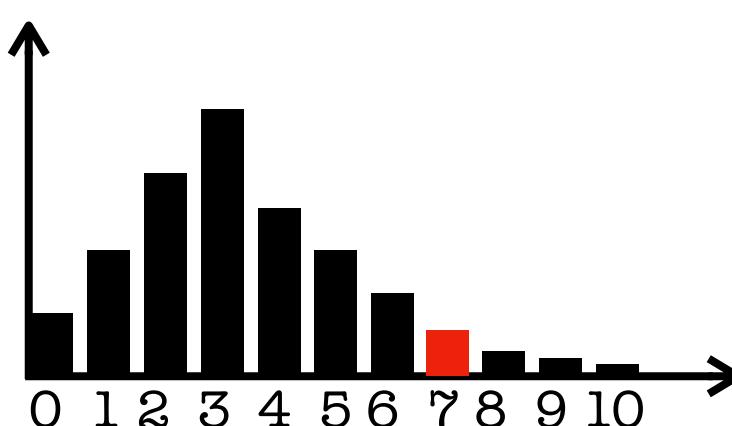
Maxima Verosimilitud: Ejemplo de lanzamiento de moneda



$$p = 0,2 \\ n = 10$$



Distribución de probabilidad para varios resultados cuando el valor del parámetro $p=0,2$ y $n=10$ lanzamientos de moneda.

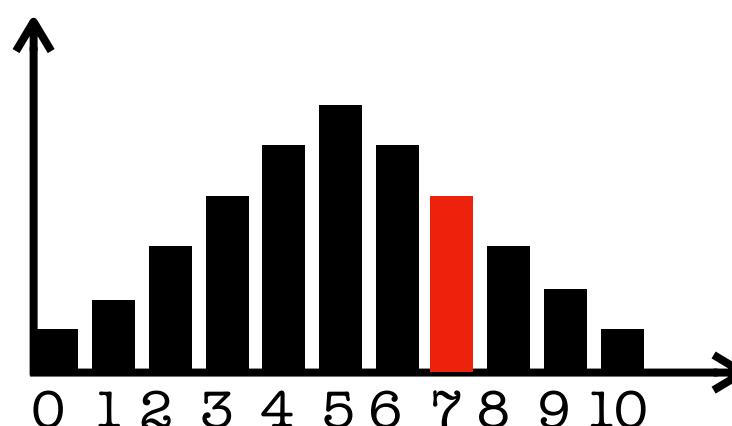


$$p = 0,3 \\ n = 10$$

Las probabilidades suman 1.

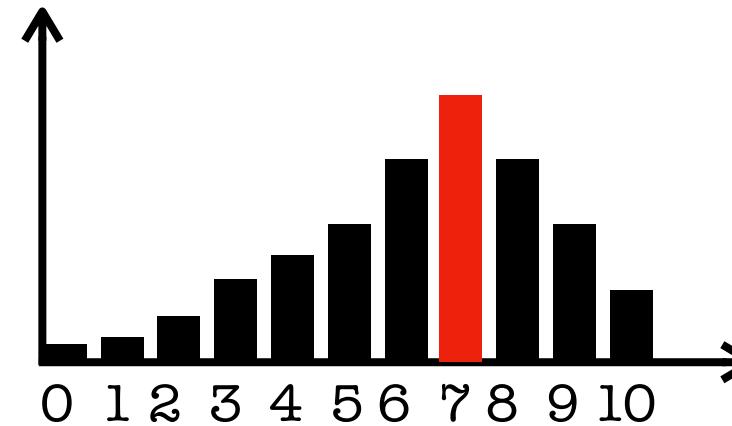
Probabilidad de que p tenga el valor 0,2 dado que observamos $x=7$ caras.

$$L_p(p=0,2 | x=7) = \Pr(x = 7 | p=0,2) = 0,001$$



$$p = 0,5 \\ n = 10$$

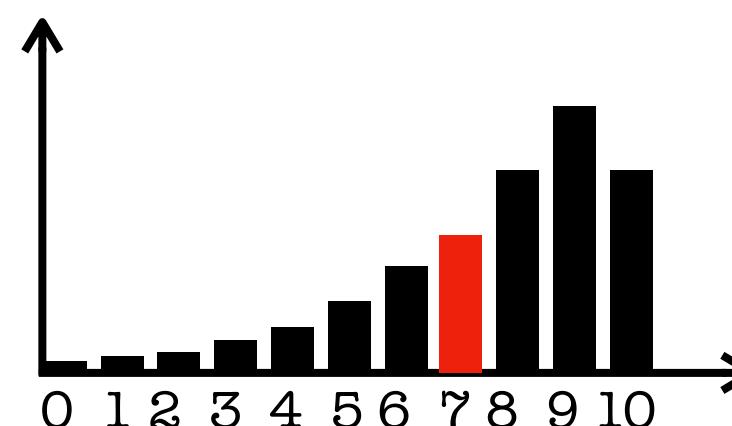
$$L_p(p=0,2 | x=7) = \Pr(x = 7 | p=0,2) = 0,001$$



$$p = 0,7 \\ n = 10$$

Probabilidad de que p tenga el valor 0,2 dado que observamos $x=7$ caras.

$$L_p(p=0,2 | x=7) = \Pr(x = 7 | p=0,2) = 0,001$$



$$p = 0,9 \\ n = 10$$

- **Datos:** resultado de lanzar una moneda 10 veces: **7 caras, 3 cruces**
- **Modelo:** la moneda tiene una probabilidad p de salir cara y $1-p$ de salir cruz

Modelo probabilístico aplicado a la inferencia filogenética

- **Datos observados:** alineamiento multiple de secuencias

H. sapiens globin A G G G A T T C A

M. musculus globin A C G G T T T - A

R. rattus globin A C G G A T T - A

- **Modelo probabilístico:**

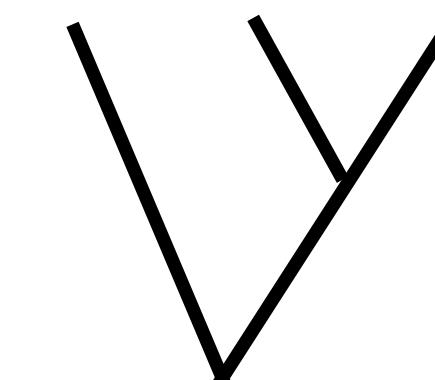
- Un modelo (hipótesis) de cómo 1 secuencia ancestral ha evolucionado en las 3 secuencias presentes en el alineamiento

- **Parámetros del modelo probabilístico (caso simple):**

- Topología del árbol y largo de las ramas
- Frecuencias nucleotídicas: π_A , π_C , π_G , π_T
- Tasas de sustitución nucleótido-nucleótido (o probabilidades de sustitución)

	A	C	G	T
A	-3 α	α	α	α
C	α	-3 α	α	α
G	α	α	-3 α	α
T	α	α	α	-3 α

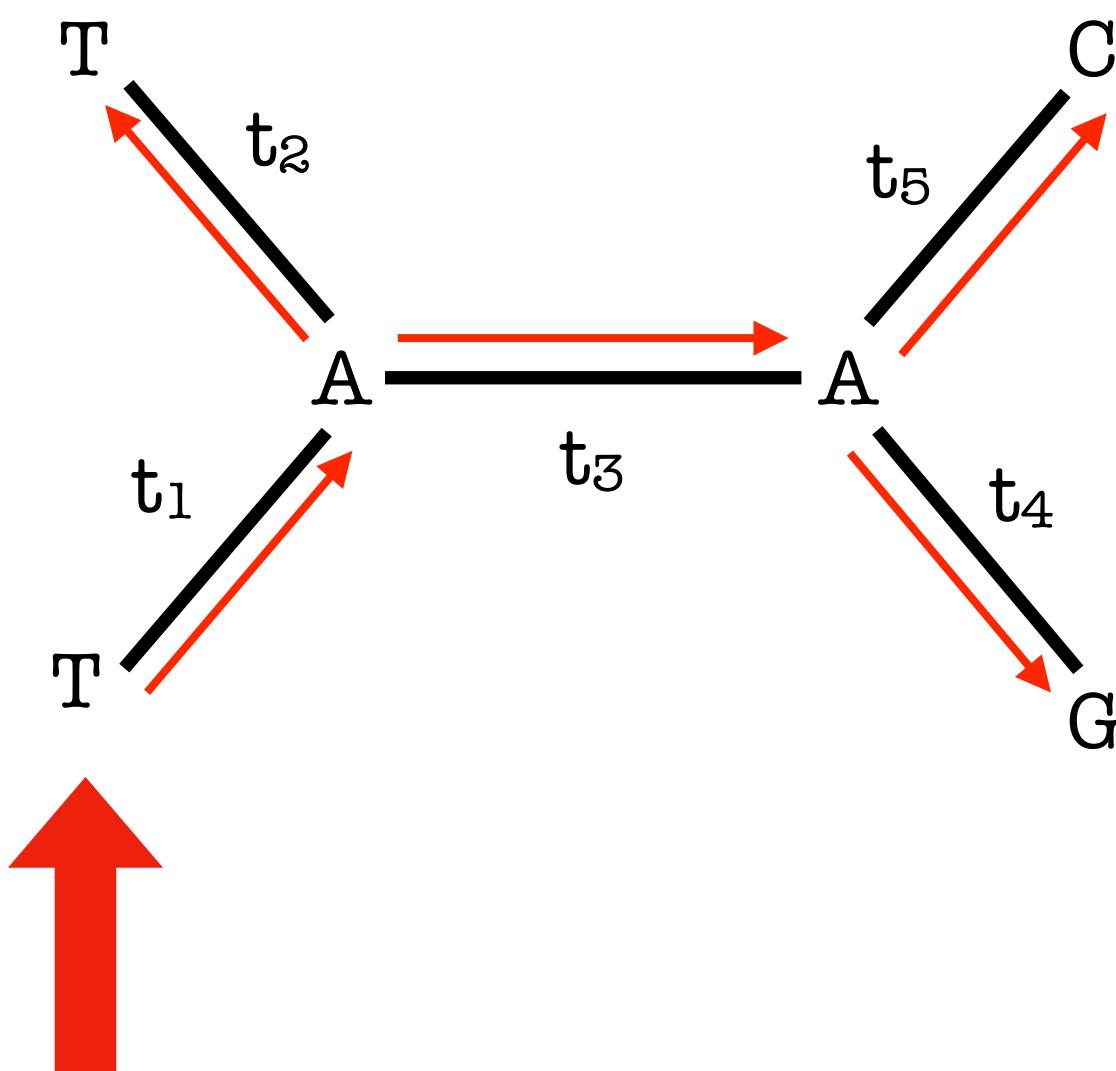
$$\Rightarrow P(t) = e^{Qt} = \begin{pmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{pmatrix}$$



Cálculo de la verosimilitud para un alineamiento dada la topología del árbol y otros parámetros.

A	T	G	G	A	T	T	C	A
A	T	G	G	T	T	T	-	A
A	C	G	G	A	T	T	-	A
A	G	G	G	T	T	T	-	A

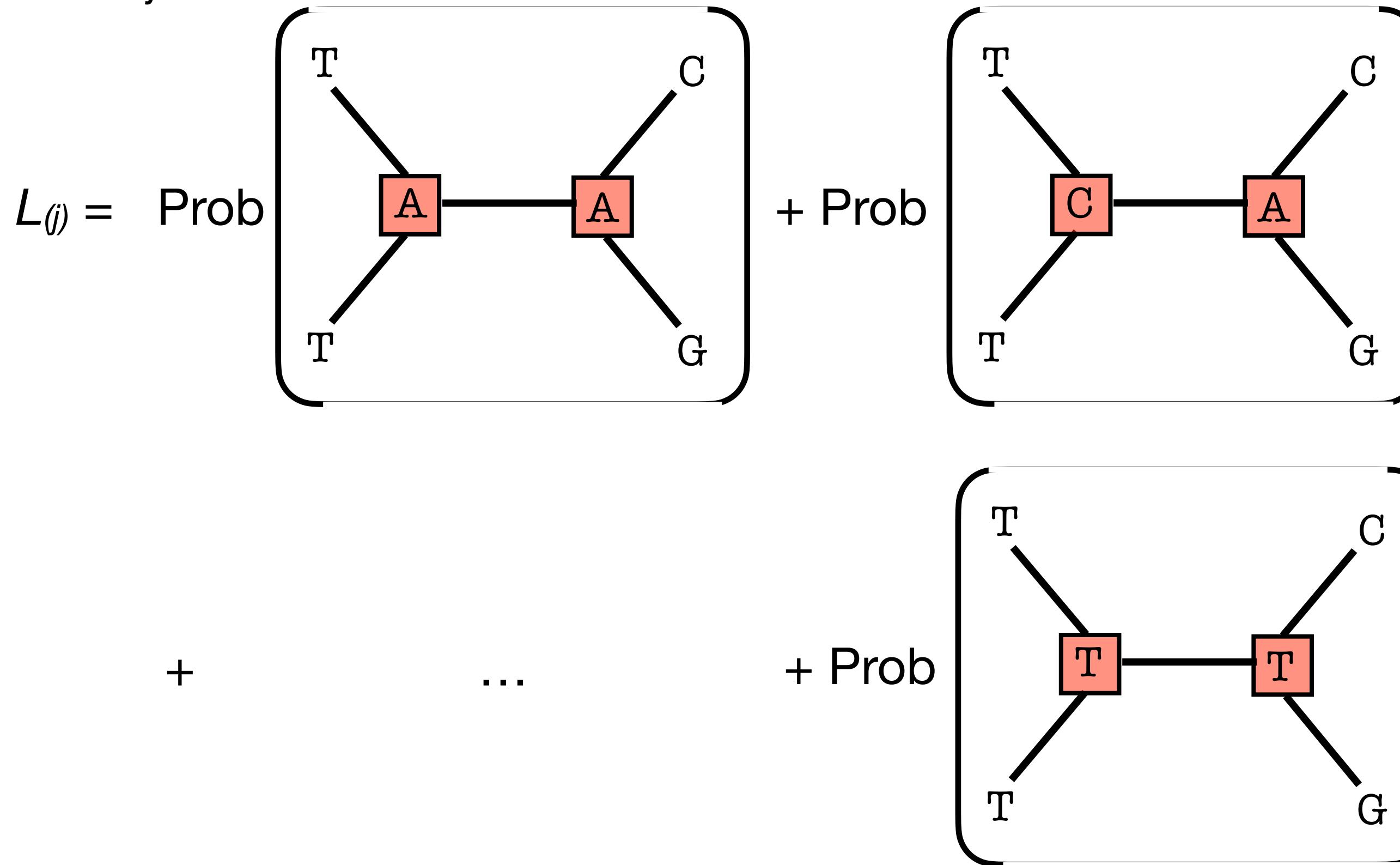
- Las columnas alineadas contienen nucleótidos homólogos.
- Asumimos la topología del árbol, las longitudes de las ramas y otros parámetros.
- Por ahora, supongamos que los estados ancestrales eran A y A (llegaremos al cálculo completo en la siguiente diapositiva).
- Comenzamos el cálculo en cualquier nodo interno o externo.
- Las flechas indican las «direcciones» de los cálculos («fluyendo» desde el punto de partida).



$$\text{Pr} = \pi_T P_{TA}(t_1) P_{AT}(t_2) P_{AA}(t_3) P_{AG}(t_4) P_{AC}(t_5)$$

Cálculo de la verosimilitud de una columna en una alineación dada la topología del árbol y otros parámetros.

A T G G A T T C A
 A T G G T T T - A
 A C G G A T T - A
 A G G G T T T - A
 j



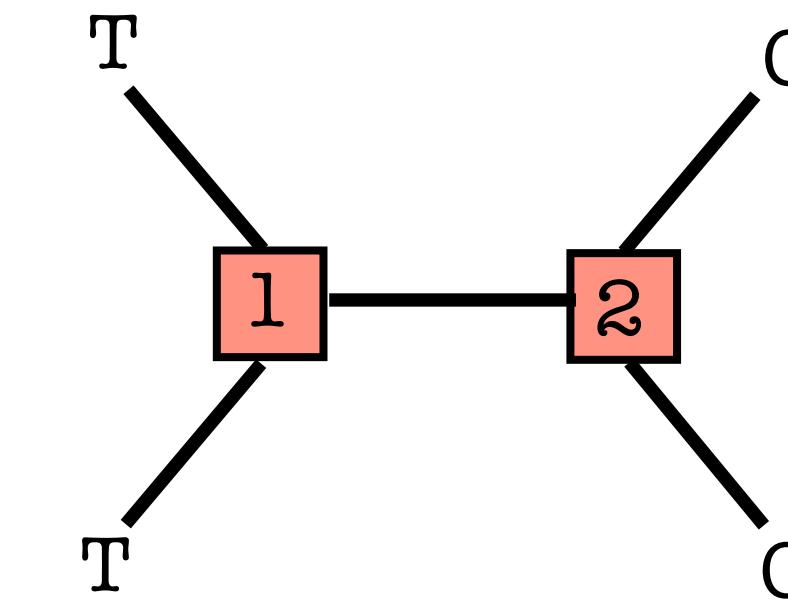
- La probabilidad debe sumarse sobre todas las combinaciones posibles de nucleótidos ancestrales.
- Aquí tenemos 2 nodos internos que dan 16 combinaciones posibles.
- Las probabilidades de las columnas individuales se multiplican para obtener la probabilidad global del alineamiento, es decir, la verosimilitud del modelo.
- En los programas de filogenia, estos cálculos se realizan mediante la suma de los logaritmos de las probabilidades (“*log likelihoods*”), ya que la multiplicación de un gran número de términos de probabilidad puede provocar un desbordamiento por debajo del rango (problema informático causado por números muy pequeños).

$$L = L_{(1)} * L_{(2)} * \dots * L_{(N)} = \prod_{j=1}^N L_{(j)}$$

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

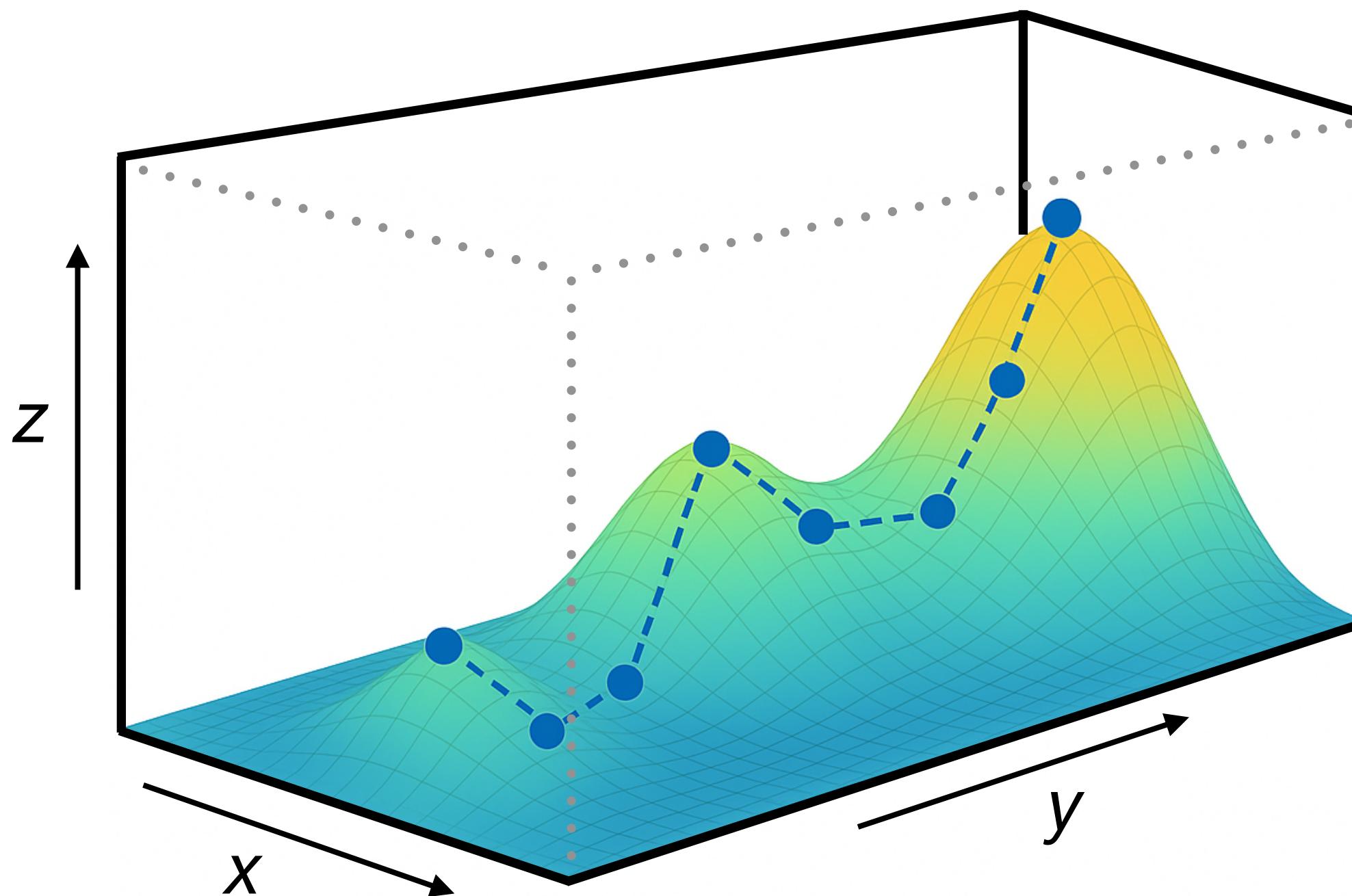
Verosimilitud de una columna en el alineamiento: calcular para cada par posible de nucleótidos ancestrales

Nodo 1	Nodo 2	Verosimilitud
A	A	0,0000009
A	C	0,0000009
A	G	0,0000009
A	T	0,0000000
C	A	0,0000001
C	C	0,0000141
C	G	0,0000014
C	T	0,0000000
G	A	0,0000001
G	C	0,0000018
G	G	0,0000150
G	T	0,0000001
T	A	0,0000248
T	C	0,0003908
T	G	0,0004028
T	T	0,0003660
Suma		0,0012198

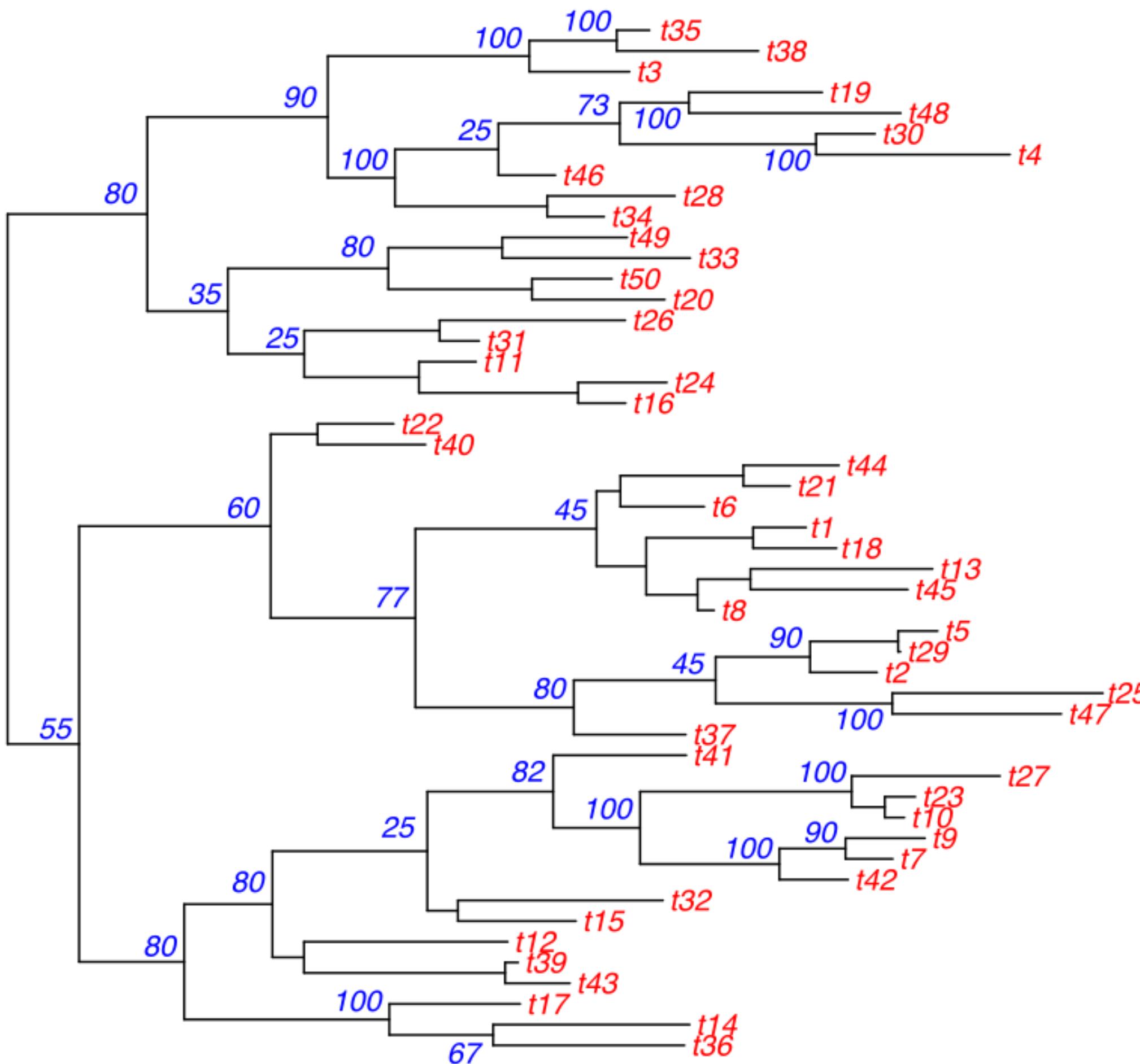


Filogenia por máxima verosimilitud

- **Datos:**
Alineamiento multiple de secuencias
- **Parámetros del modelo:**
Topología del árbol y largo de las ramas, frecuencias nucleotídicas y tasas de sustitución
 - Elija valores iniciales aleatorios para todos los parámetros, calcule la verosimilitud.
 - Cambiar ligeramente los valores de los parámetros en una dirección para que mejore la verosimilitud.
 - Repita hasta encontrar el máximo.
- **Resultados:**
 - Estimación ML de la topología del árbol
 - Estimación ML de la longitud de las ramas
 - Estimación ML de otros parámetros del modelo
 - Medida de lo bien que el modelo se ajusta a los datos (verosimilitud)



Bootstrap: es una forma de comprobar la fiabilidad de las ramas (o agrupaciones) del árbol



Cómo funciona

- 1) Alineamiento
 - 2) Muestrear posiciones con reemplazo para crear una pseudo-alineamiento
 - 3) Construir un árbol a partir de este nuevo alineamiento
 - 4) Repetir este proceso N veces (100 a 1000 veces)
 - 5) Finalmente se comprueba para cada rama del árbol original la frecuencia en que apareció esa rama en los árboles de bootstrap

