# Construcción y análisis de árboles filogenéticos

## Bioinformática

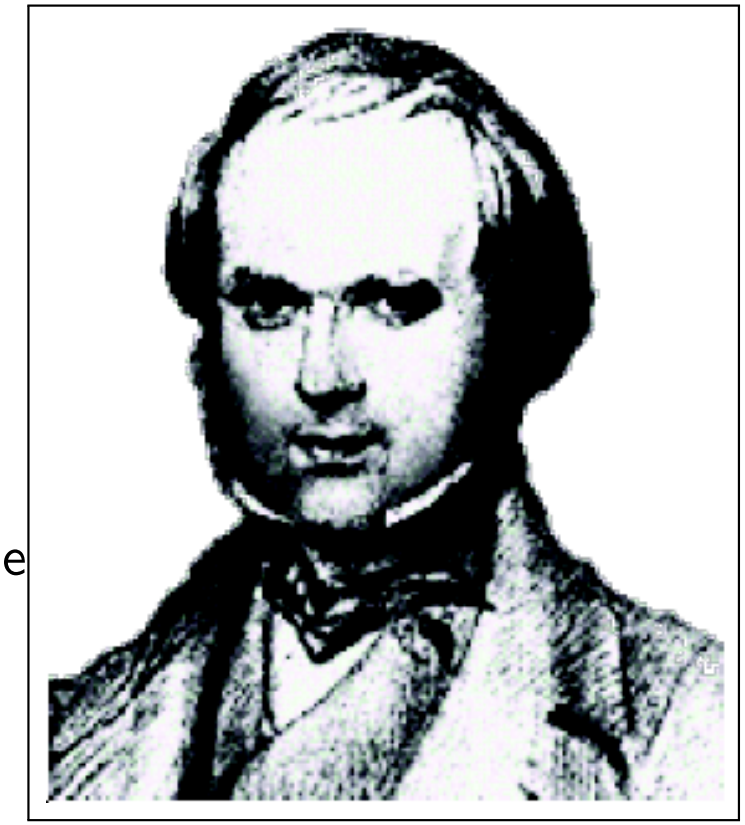
Luis E. Garreta U
luis.garreta@javerianacali.edu.co

Pontificia Universidad Javeriana – Cali
Facultad de Ingeniería - Carrera de Biología

22 de octubre de 2018

# Filogenía

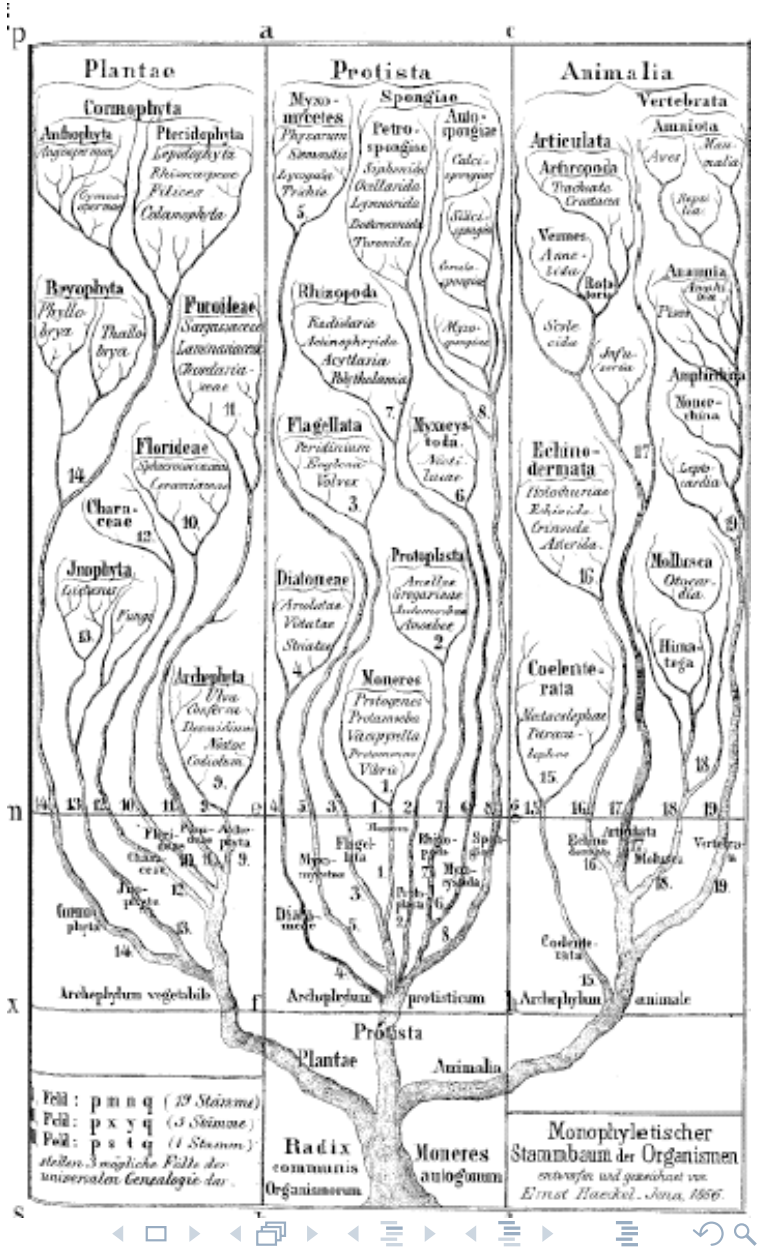
- La filogenia es el estudio de las relaciones evolutivas.
- Surge a partir de la teoría de la evolución de Darwin.
- Son representaciones gráficas de las relaciones evolutivas entre un grupo de organismos vivos.



# Árboles Filogenéticos

Son representaciones gráficas de las relaciones evolutivas entre un grupo de organismos vivos:

- Primer árbol filogenético debido a Haeckel 1866
- Todas las especies descienden por evolución de una especie ancestral común
- La aparición de una nueva especie se produce por la subdivisión de una existente en dos subespecies que han divergido tanto que pierden la capacidad de cruzarse.



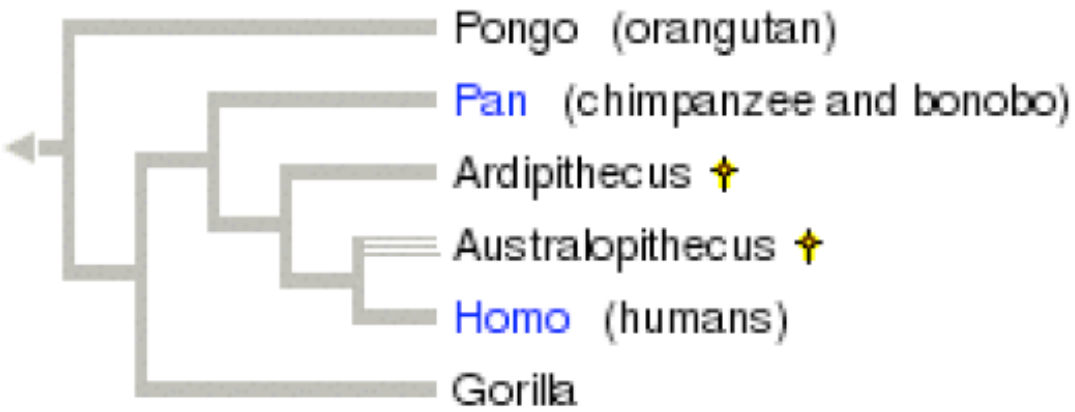
# El árbol de la vida

# Qué muestran las Filogenías

- Un análisis filogenético no sólo nos indica:
  - ✓ Relaciones evolutivas entre las secuencias o especies,
  - ✓ *Es decir, cuales* descienden de ancestros comunes
- También puede indicarnos cuales son las distancias entre ellas.

# Algo más cercano

# Algo más cercano

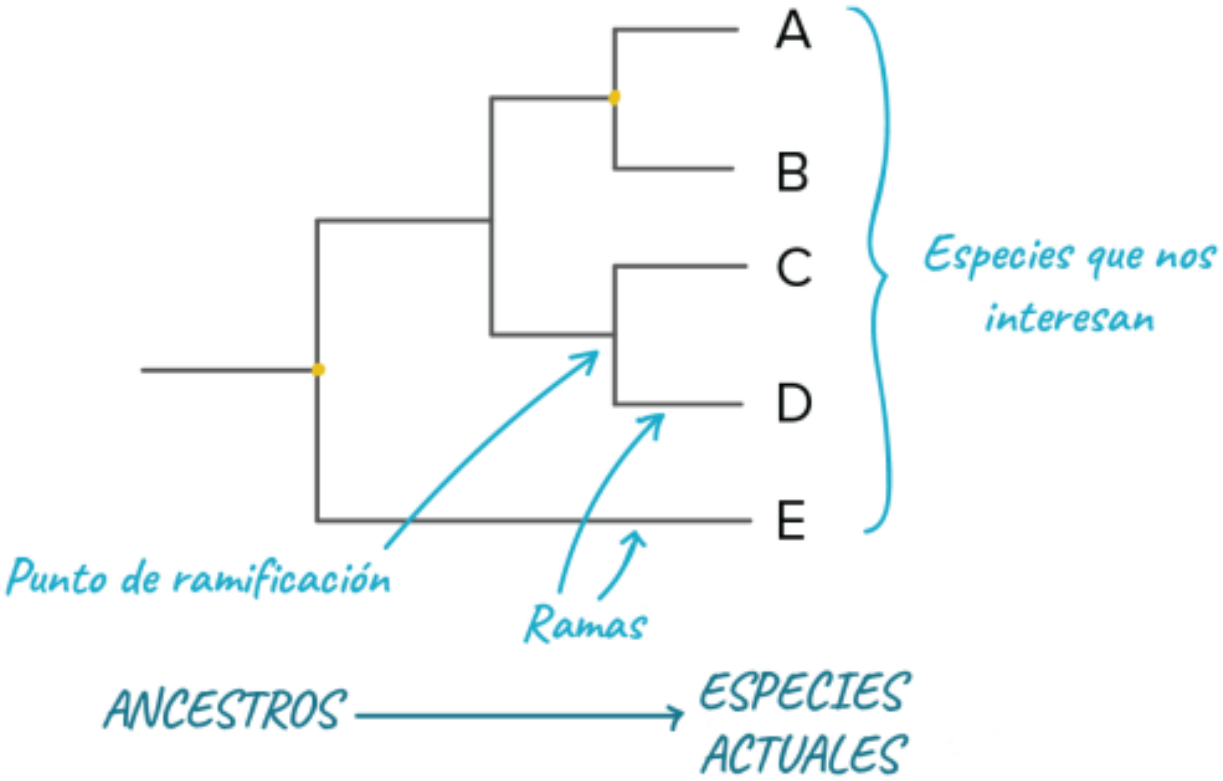


# Qué se asumen en las Filogenias

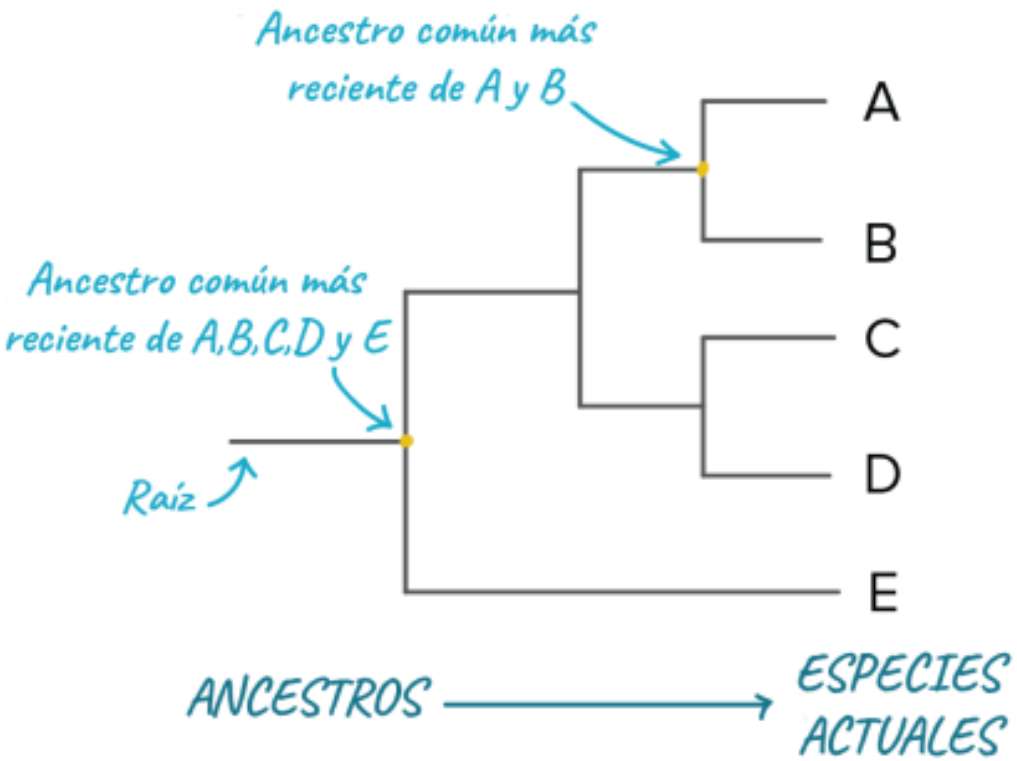
- Asumen que todas las secuencias o especies de las que tenemos información son **especies actuales** y que ninguna de **ellas es un antepasado de cualquiera de las otras**.

- Los métodos de reconstrucción filogenética más habituales asumen que todas las secuencias o especies **provienen de partir un ancestro común** mediante bifurcaciones.

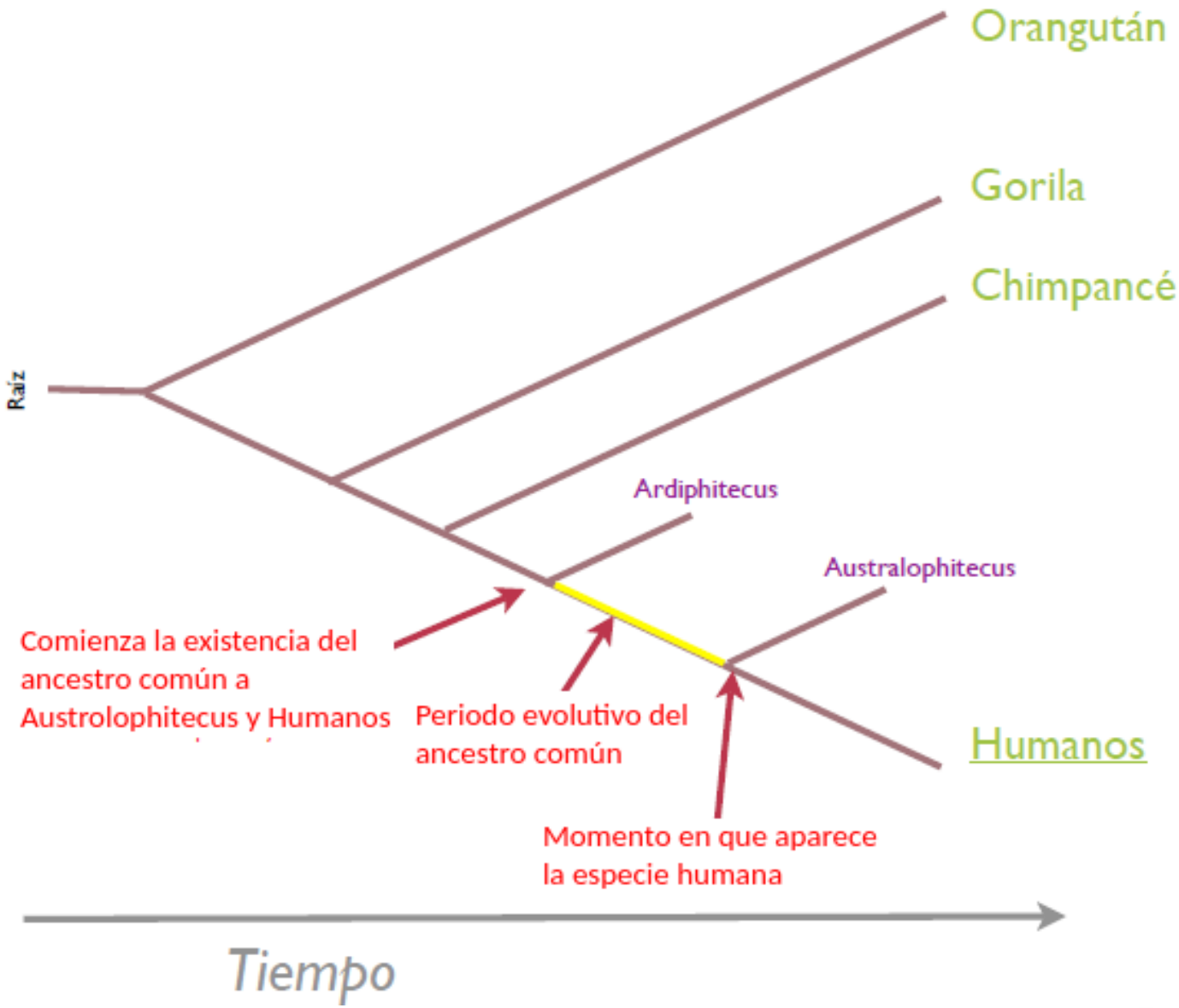
# Conceptos Árboles Filogéneticos: Especies, ramas, nodos



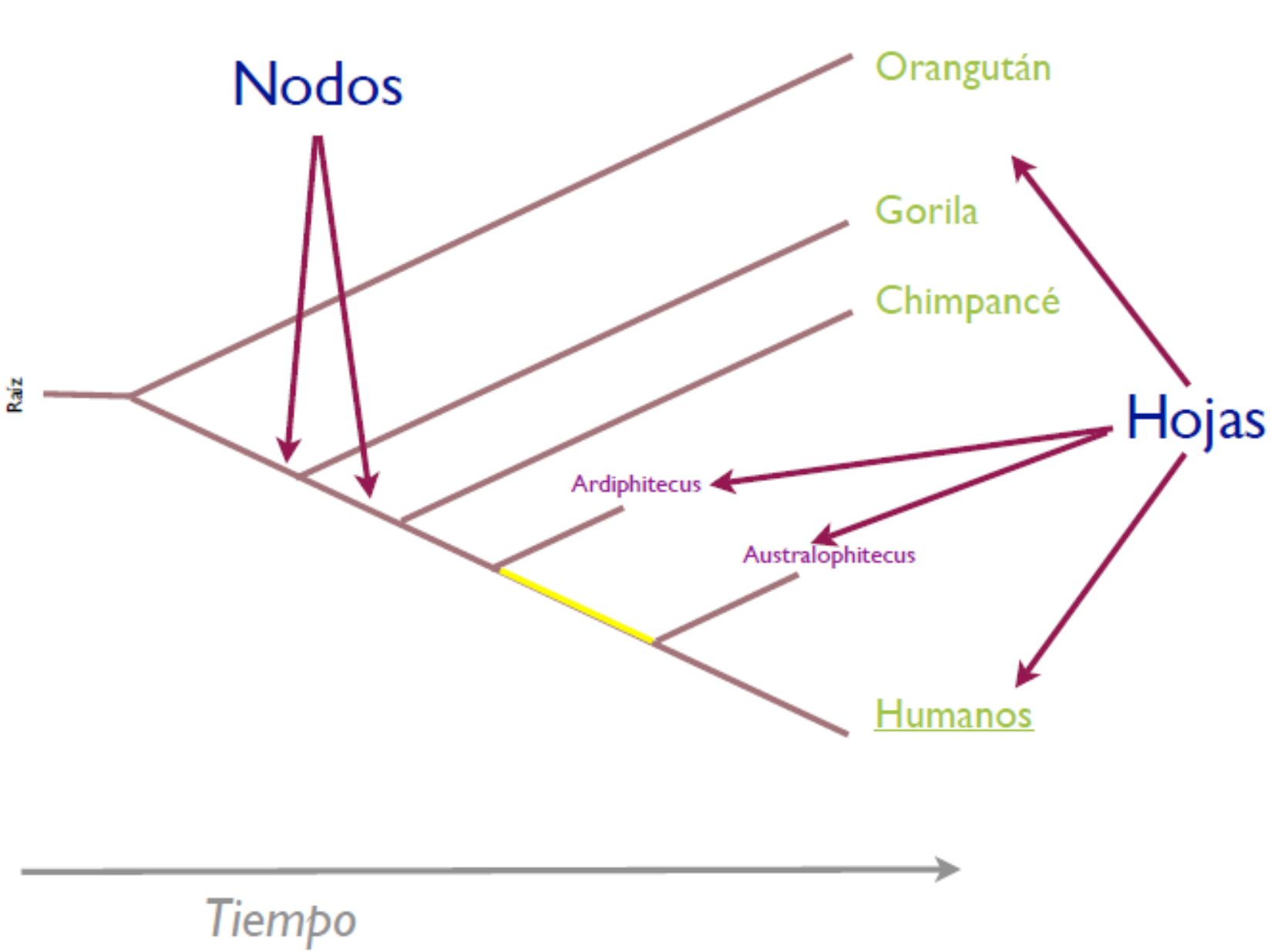
# Conceptos Árboles: ancestros



# Ejemplo: Árbol filogenético para el grupo Hominidae

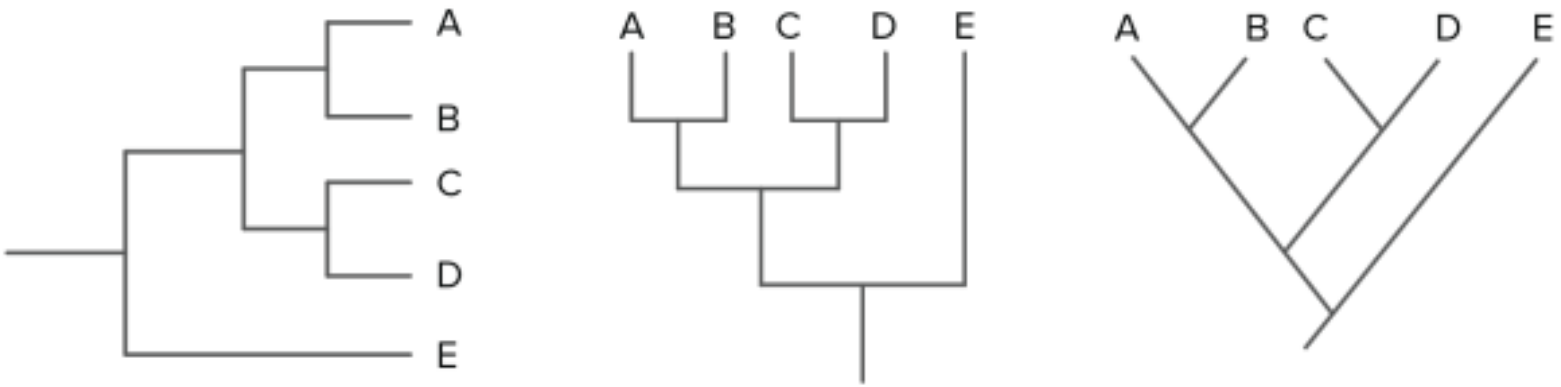


# Ejemplo: Árbol filogenético para el grupo Hominidae



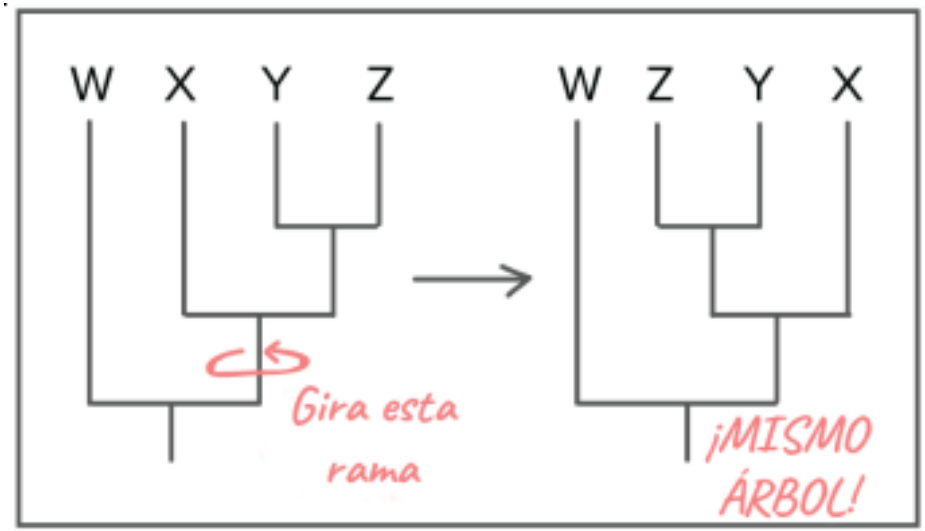
# Distintas Formas de Árboles

Relaciones idénticas mostradas en árboles diferentes



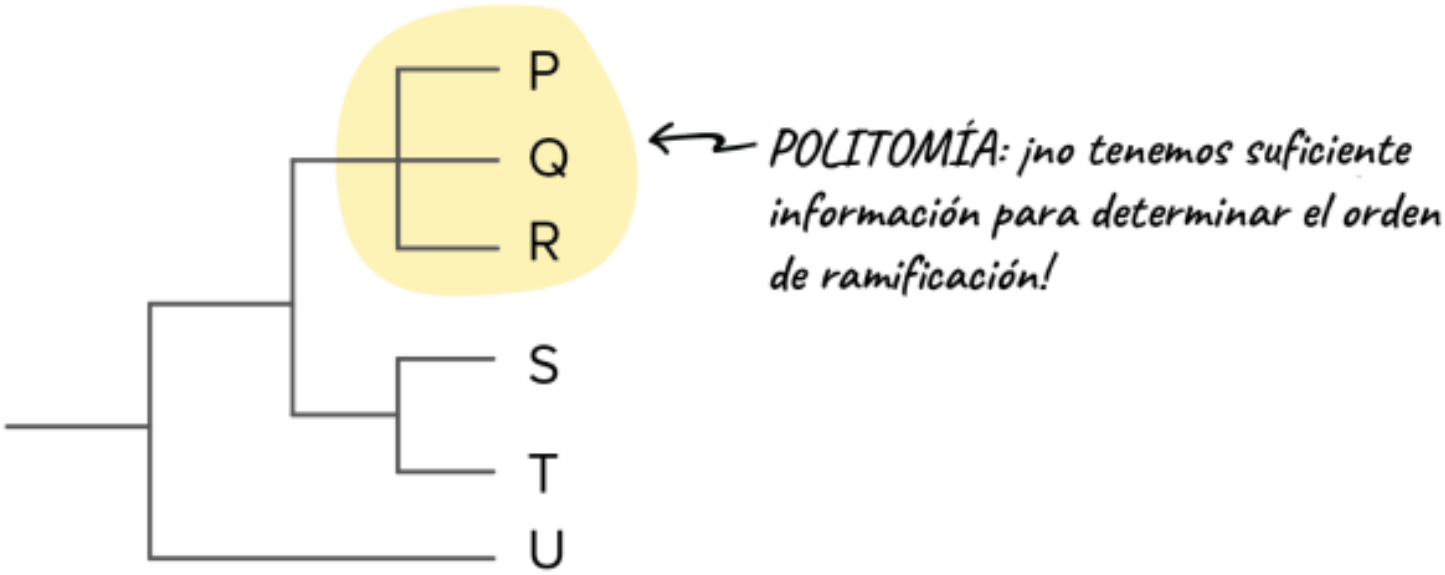
# Giros en las rámas

Algunos giros de las ramas no modifica el árbol



# Politomías

- Hasta ahora, todos los árboles tienes solo dos linajes (líneas de descendencia) que surgen de cada punto de ramificación.
- Sin embargo, puedes encontrar árboles con una politomía (poli, muchos; tomía, corte), lo que significa que un punto de ramificación tiene tres o más especies diferentes:



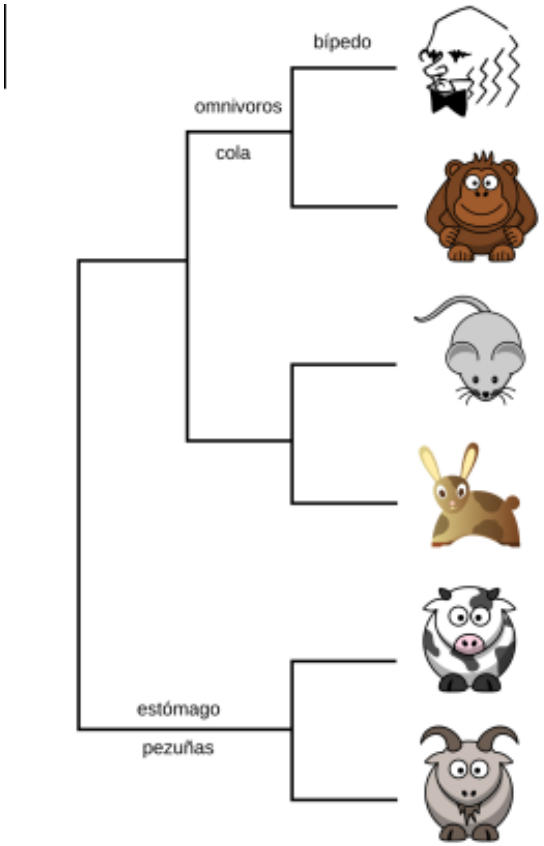
# Qúe datos se usan?

# Secuencias alineadas "sin huecos" de ADN, ARN, mARN, Proteínas

```
TGGCGCCGAG CCCGCTCCGA GCCCTGCCAG TGGCTGGCGG AGGCGACAGT
TGGCAGCGGG CACACTGAGG GGCTTGGCAG TGGCTGGTGG AGGCGAGAGC
TGGCGGCGGG CACGCTGAGG GGCTTGGCAG TGGCTGGTGG AGGCGAGAGC
TGGCGCCGGG AGCGCTACCC GCCCTGCCAG TGGCCGGCGG AGGCGACAGT
TGGCGGCGAG CGCGCTGCGA GGCCTGCCAG TGGCTGGCGG AGGCGACAGT

AGCGACAGCG AGGATGACCG CTGCGAGATT GGGTATCTCG ACCGGACGTC
AGCGATAGTG AGGATGACGG CTGGGATATT GGGTATCTCG ACCGGTCCTC
AGCGATAGTG AGGATGACGG CTGGGATATT GGGTATCTCG ACCGGTCCTC
AGCGACAGCG AGGATGACCG CTGCGAGATT GGGTATCTCG ACCGGACGTC
AGCGAGAGCG AGGATGATGG CTGGGAGATT GGGTATCTCG ACCGGACGTC

TCAGAAATTG AAAGGGCTGT TACCCATTGA AGAAAAGAAA GAAAAATTTA
TCAGAAATTA AAAAGGTCTT TACCCGTTGA AGAGAAGACC GAGACATTTA
TCAGAAATTA AAAAGGTCTT TACCCGTTGA AGAGAAGAAC GAAACATTCA
TCAGAAATTG AAAGGGCTGT TACCCATTGA GGAAAAGAAA GAAAAATTTA
TCAGAAATTG AAAAGGCTAT TACCCATTGA AGAAAAGAAA GAAAAATTTA

AGAAAGCAAT GACCATCGGA GATGTTTCAT TGGTCCAGGA GCTCCTAGAT
AGAAAGCACT GACCACTGGA GATATTTCTT TAGTGAAAGA ACTCCTGGAT
AGAAAGCACT GACCACCGGA GATATTTCCT TAGTGAAGGA ACTCCTGGAT
AGAAAGCAAT GACCATCGGA GATGTTTCAT TGGTCCAGGA GCTCCTAGAT
AGAAAGCAAT GACCATCGCA GATCTTTCAT TGGTCCAGGA GCTCCTAGAT

TCTGGCATTA GTGTAGATTC CACCTTTCAG TATGGATGGA CTCCCCTTAT
TCTGGCATTA ATGTGGACTC CAGCTTCCGC TATGGATGGA CCCCTCTTAT
TCTGGCATTA ATGTAGATTC CAGCTTCCGC TATGGATGGA CCCCTCTTAT
TCTGGCATTA GTGTAGATTC CACCTTTCAG TATGGATGGA CTCCCCTTAT
TCTGGCATTA GTGTAGATTC CAACTTTCAG TATGGATGGA CTCCCCTTAT

GTATGCTGCT AGTGTTGCCA ATGCAGAGCT GGTTCGGGTC CTTTTGGACA
GTATGCCGCT AGTGTGCCCA ATGCAGAGCT GGTTCGCTTC CTTTTGGACA
GTATGCAGCT AGTGTGGCCA ATGTAGAGCT GGTTCGGTTC CTTTTGGACA
GTATGCTGCT AGTGTTGCCA ATGCAGAACT GGTTCGGGTC CTTCTGGACA
GTATGCTGCT AGTGTTGCCA ATGCAGAGCT GGTTCGGGTC CTTTTGGACA
```

# Tabla de caracteres morfológicos codificados

| | Alimentación | Estómago | Pezuñas | Bípedo | Cola |
|---|---|---|---|---|---|
| Hombre | Omnívoro | Simple | No | Sí | No |
| Chimpancé | Omnívoro | Simple | No | No | No |
| Ratón | Herbivoro | Simple | No | No | Sí |
| Conejo | Herbivoro | Simple | No | No | Sí |
| Vaca | Herbivoro | Compuesto | Sí | No | Sí |
| Cabra | Herbivoro | Compuesto | Sí | No | Sí |



# Otros Datos

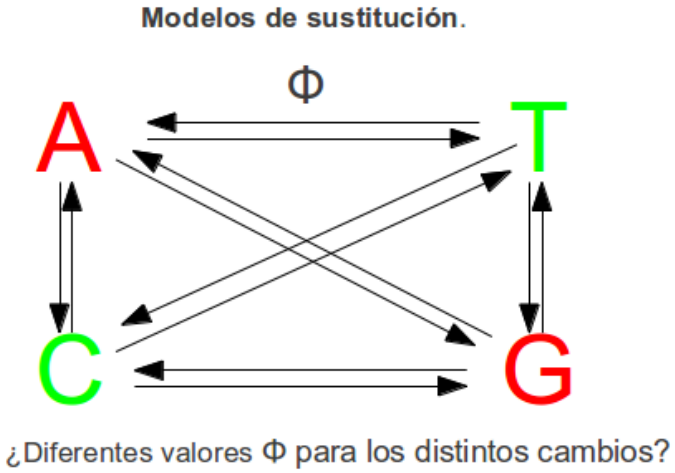
- Lista ordenada de genes si se dispone del genoma completo
- Lugares de restricción, SNPs, Secuencias de aminoácidos, etc

# Análisis Filogenético

# Fases del Análisis Filogenético

- Selección de las secuencias a analizar
  - ✓ A partir de una de las Bases de Datos vistas, en formato fasta
- Análisis múltiple de secuencias
  - ✓ Mediante uno de los métodos o herramientas vistas
- Elección de un modelo de sustitución
- Construcción del árbol (inferencia filogenética)
- Evaluación del árbol

# Modelos de sustitución




- Podemos calcular las distancias entre las secuencias asumiendo distintos modelos de mutación
- Hay modelos más sencillos y otros más complejos
- Según los datos de los que dispongamos el modelo de mutación óptimo puede ser uno u otro.
- Existen programas, como jModelTest, que pueden probar todos estos modelos en nuestros datos y nos pueden recomendar qué modelo debemos utilizar para realizar un árbol de forma óptima.

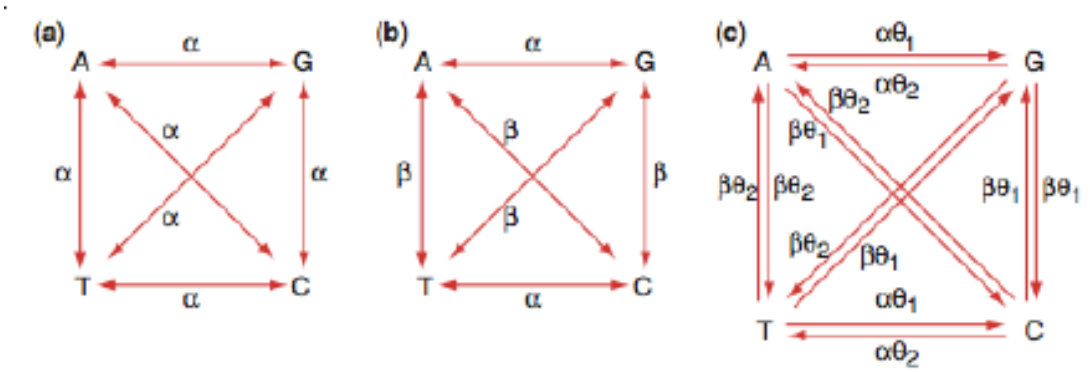
# Varios Modelos de Sustitución

- Modelo de Jukes-Cantor
  - ✓ Considera que la probabilidad de sustitución es igual para todas las combinaciones de nucleótidos/aminoácidos

$d = -3/4 \ln (1-4/3p)$
p proporción de cambios entre las dos secuencias

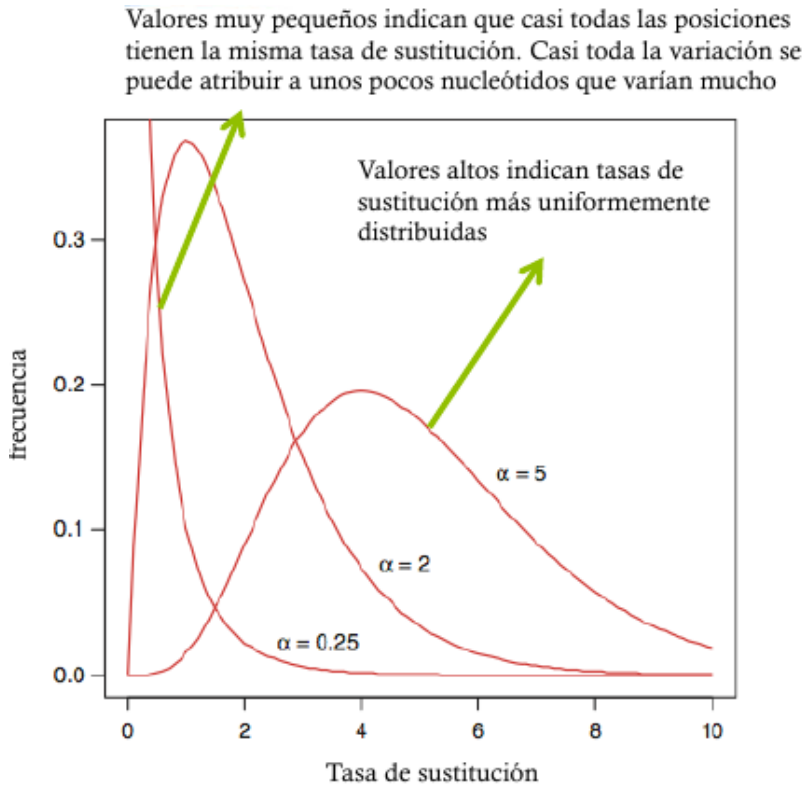
- Modelo Kimura2-parametros:
  - ✓ Distingue entre transiciones (purina a purina o pirimidina a pirimidina) y transversiones (purina a pirimidina y viceversa)

$d = -1/2 \ln(1-2p-q) - 1/4 \ln(1-2q)$
p proporción de transiciones y q proporción de transversiones.



# Modelo General

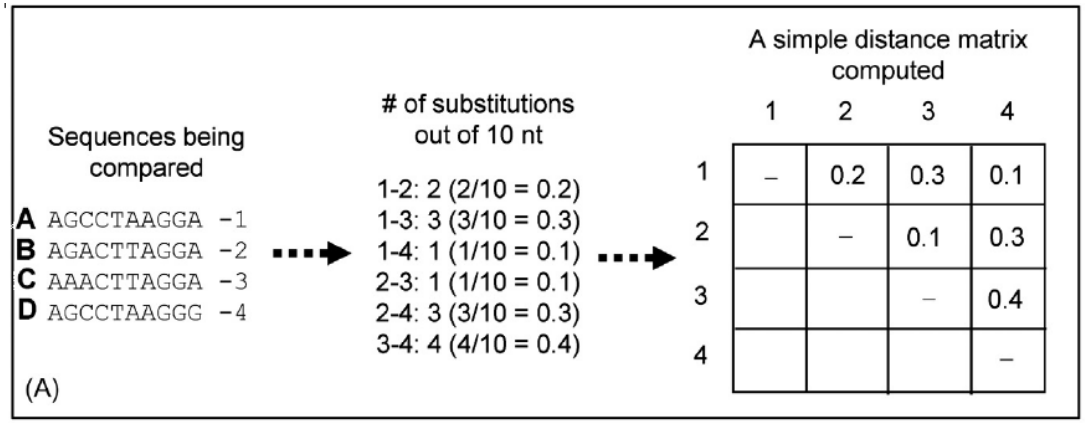
- Algunas posiciones dentro de la proteína varían mucho y otras muy poco (Distribución Gamma):
  - ✓ La tercera posición de un codón suele tener una tasa de sustitución más alta que los dos primeros (código degenerado)
  - ✓ Algunas regiones de las proteínas tienen dominios conservados
- Para ello se asocia una tasa de sustitución distinta a cada posición, usando una distribución gamma
  - ✓ El parámetro $\alpha$ modula la forma de la distribución
  - ✓ Proteínas que evolucionan rápidamente tienen una $\alpha$ pequeña

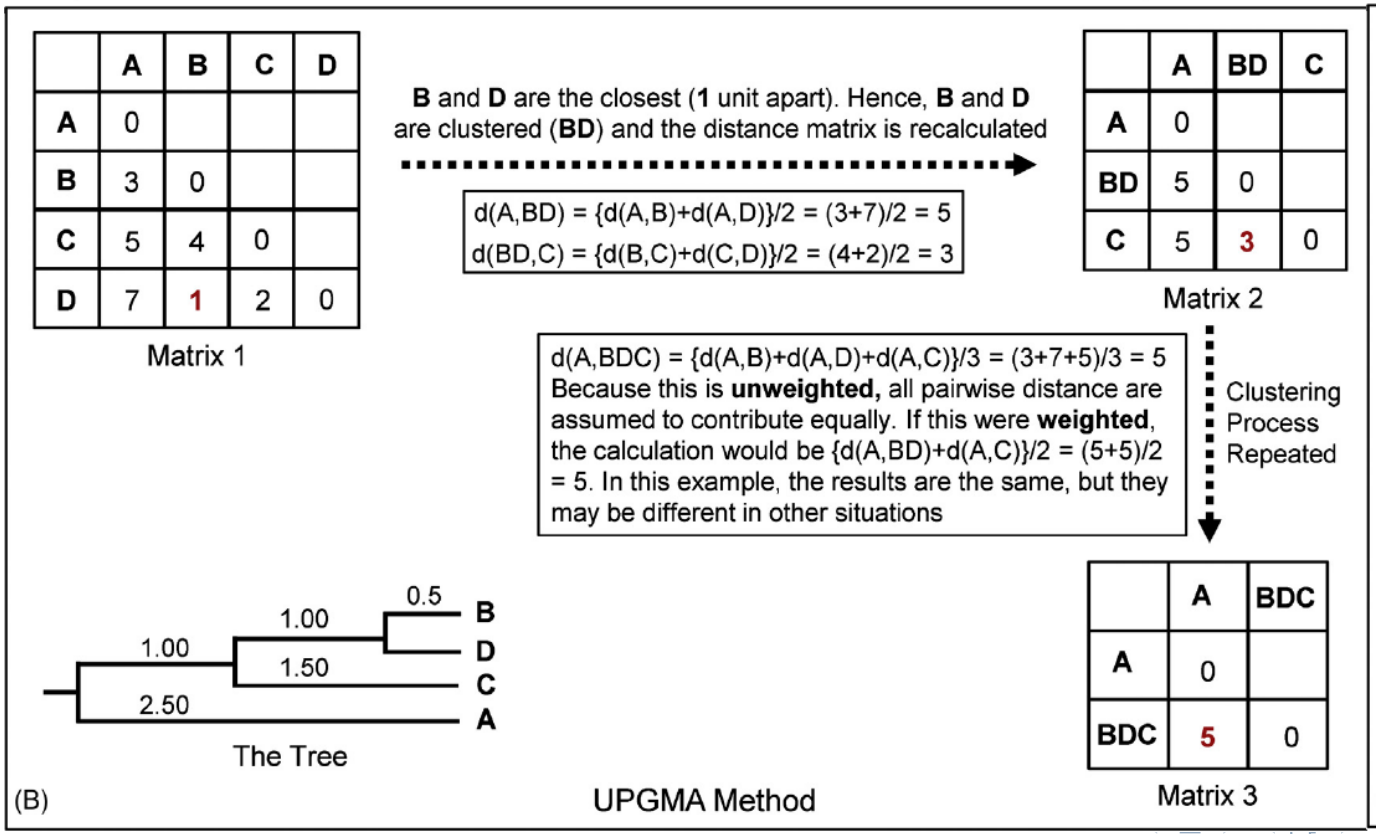


# Métodos de Construcción de Árboles

- **Métodos de distancia**:
  - ✓ Construye el árbol basado en la distancia evolutiva para todas las OTUs
- **Máxima Parsimonia**:
  - ✓ Construye un árbol que minimice el número de cambios requeridos para explicar los datos
- **Máxima Similitud**:
  - ✓ Construye un árbol que maximiza la probabilidad de ser el generador de las secuencias observadas.
- **Bayesianos:**
  - ✓ Calcula una probabilidad posterior para cada árbol posible dado un modelo de evolución y unas observaciones.

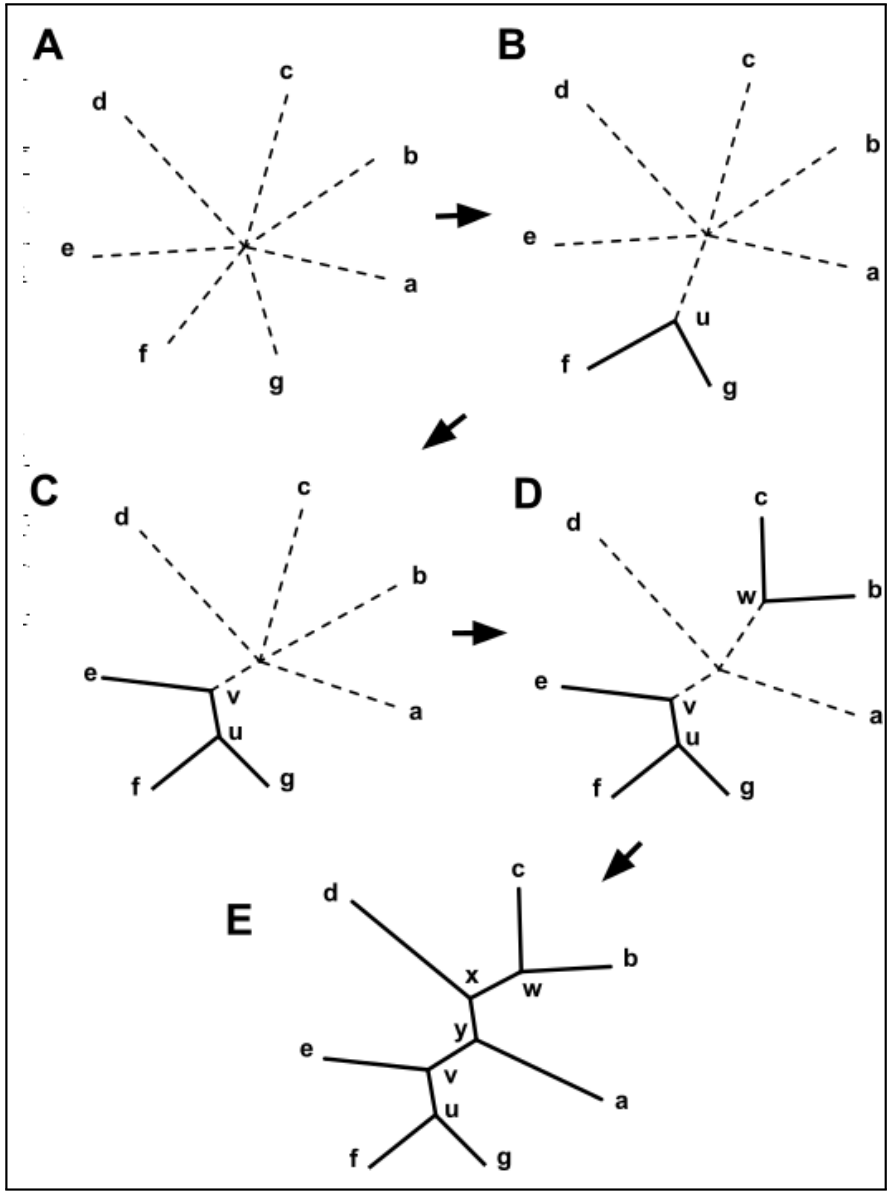
# Método de Distancias



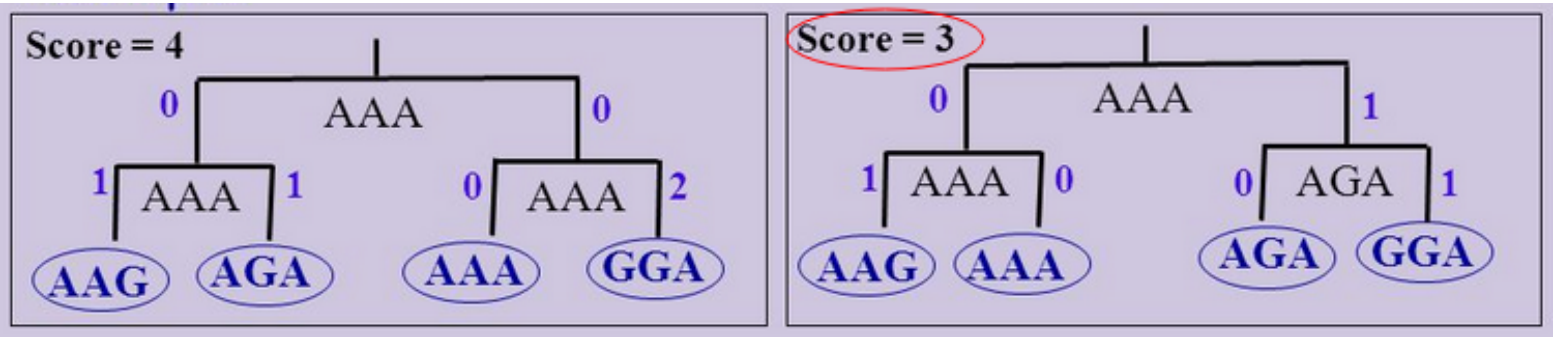
# Algoritmos de Métodos de Distancias

- Agrupamiento UPGMA:
  - ✓ Se basa en agrupar los pares con distancias más cercanos
- Agrupamiento de **Neighbor Joining:**
  - ✓ Lo mismo que el anterior, pero tenga en cuenta también las distancias a las otros nodos



# Máxima parsimonia

- El método de máxima parsimonia se basa en la filosofía de que la explicación más simple, la que requiere menos cambios debe ser la correcta. Mediante este método se obtienen árboles que ordenan las ramas de modo tal que se minimiza el número de mutaciones que deben haber ocurrido.
- Ejemplo:




- Para elegir el mejor árbol, el árbol que implicase menos cambios, en teoría habría que evaluar todos los árboles posibles.

# Máxima verosimilitud

- El método de máxima verosimilitud busca el árbol máximoverosímil, es decir, el árbol que es más probable que haya generado los datos que hemos observado.
  - ✓ Evidencias + modelo -> árbol más probable
- Elección modelo mutación:
  - ✓ más parámetros, mejor ajuste, peores varianzas.
  - ✓ programas para quedarse con el más adecuado a nuestros datos.
- **A partir de los mismos datos mejor estadísticamente que distancias y parsimonia.**
  - ✓ No problema con ramas largas si hay suficiente información
- Desventajas:
  - ✓ Coste computacional.

# Evaluación de Árboles

# Evaluación de los árboles

- Que un programa informático produzca un árbol filogenético no significa que sea correcto:
  - ✓ GIGO (Garbage In, Garbage Out)
- En muchos casos puede ser globalmente correcto pero tener inexactitudes en algunas ramas
- Evaluación: bootstrapping o remuestreo:
  - ✓ Verificación del significado biológico de un árbol evaluando su robustez

# Bootstraping

- Primero, seleccionamos columnas del AMS original de forma aleatoria, hasta tener tantas como en el AMS original
  - ✓ Se permiten repeticiones (muestreo con reemplazamiento)
  - ✓ Es un alineamiento artificial, pero que conserva las características del AMS original
  - ✓ Se realizan muchos de estos muestreos aleatorios (100 a 1000)
- A cada AMS aleatorio se le aplica el algoritmo a evaluar, obteniendo un árbol
- Se construye un árbol de consenso con todos los árboles obtenidos:
  - ✓ El porcentaje de veces que una ramificación aparece es el valor de bootstrap
  - ✓ Valores de bootstrap $> 70\,\%$ suelen tomarse como suficientemente robustos (equivalen a un nivel de significatividad $p < 0.05$)

# Idea Bootstraping