Introducción a la Bioinformática

Computational Phylogenetics

Filogenias, Reconstrucción filogenética, Inferencia de filogenias

Fernán Agüero

Instituto de Investigaciones Biotecnológicas, UNSAM

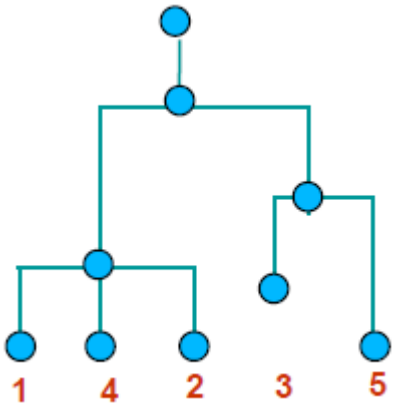
Filogenia

- Una filogenia es un árbol que describe la secuencia de eventos que llevó a producir los caracteres que observamos en la actualidad
- Es una hipótesis!
- Los eventos pasados son desconocidos. Se *infieren*

- Un árbol es un grafo
  - Nodos y ejes
- En particular:
  - Los nodos exteriores (hojas del árbol) son los eventos observados (especies actuales)
  - Los nodos internos son los eventos (ancestros) postulados
  - La longitud de los ejes (ramas) representa el tiempo de evolucion e nodos

Ancestral sequences →

Input sequences →



Computational phylogenetics

Es la aplicación de algoritmos computacionales, métodos y programas, al **análisis filogenético.**

Análisis filogenético = inferencia de una hipótesis que explique las relaciones ancestrales (de herencia).

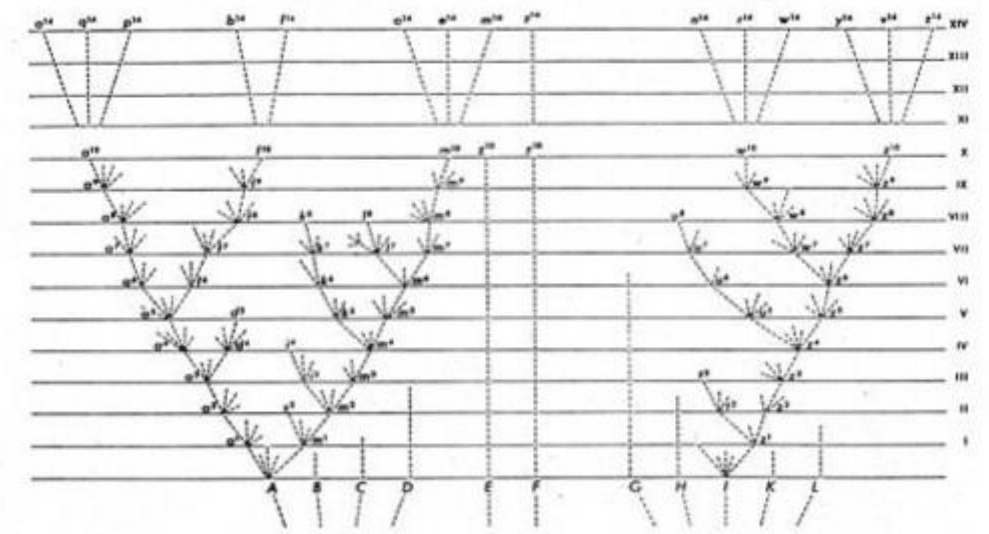
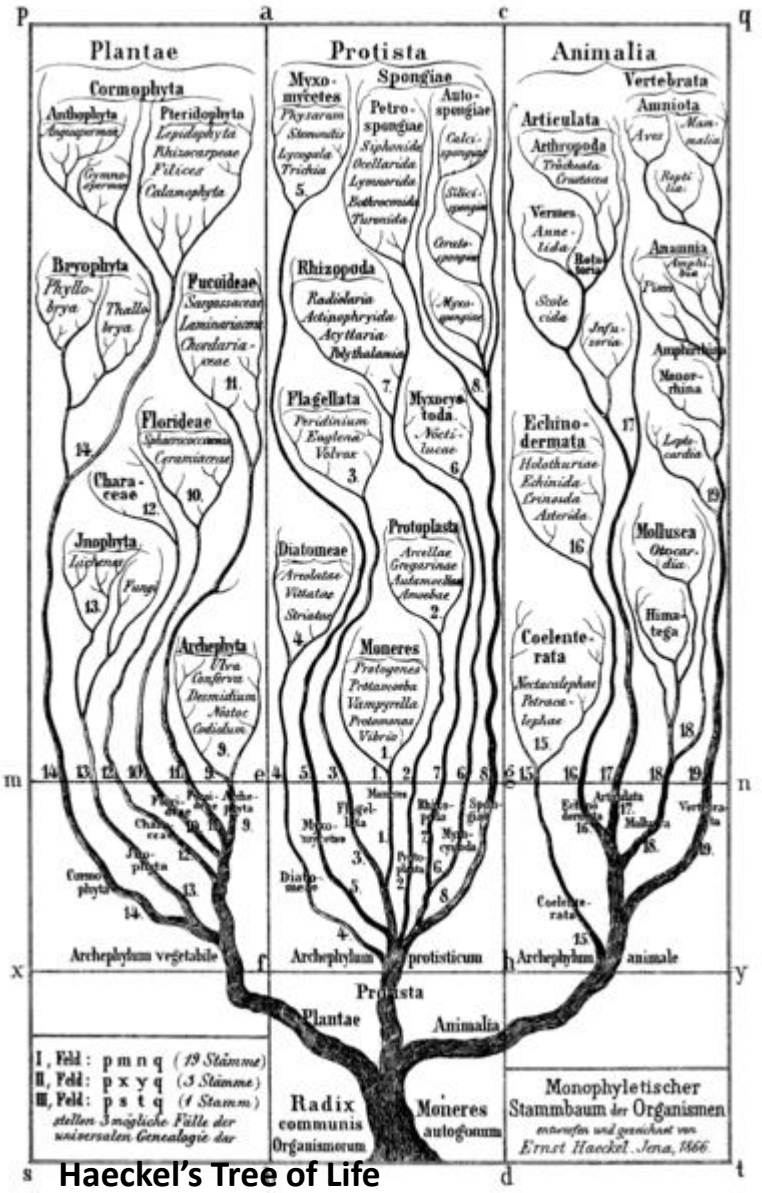


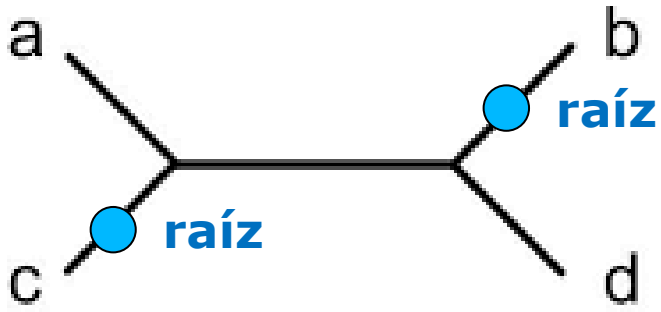
Diagrama de Darwin sobre el Origen de las Especies

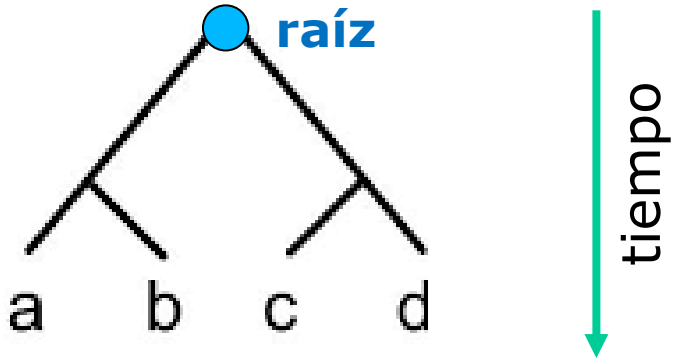


Haeckel's Tree of Life

Tipos de árboles

Arboles con raíz (rooted) vs Arboles sin raíz (unrooted)





Los dos árboles explican las relaciones filogenéticas.

El árbol con raíz además muestra la ***dirección*** de la evolución.

Poniendo el problema en perspectiva computacional

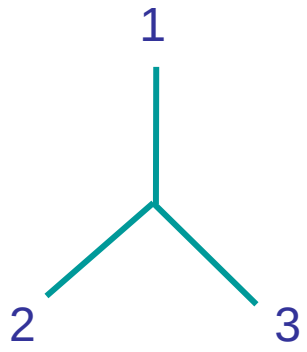
Espacio de árboles posibles

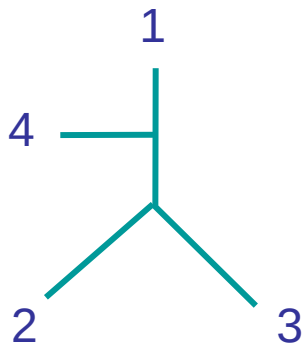
Taxa	Rooted	Unrooted
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
80	2.18 E+137	3.43E+139

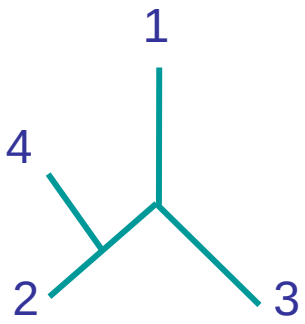
Brute Force Approach:

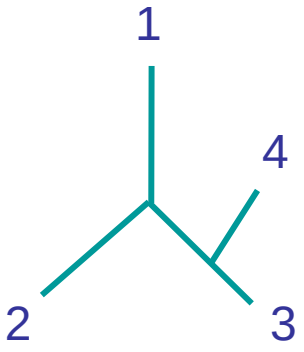
1. Build all trees
2. Calculate a score for each tree
3. Identify the best tree

3'. Identify the best trees, build a consensus tree









Poniendo el problema en perspectiva

Para *n* taxones (nodos), el número de topologías diferentes (árboles) es:

Rooted: $$\frac{(2n-3)!}{2^{n-2}(n-2)!}, \text{for} \, n \geq 2$$

Unrooted: $$\frac{(2n-5)!}{2^{n-3}(n-3)!}, \text{for} \, n \geq 3$$

- **Basados en**
  - **Distancias**
    - Distancia utilizada = genética
    - Estrategia computacional = Data clustering
    - Usa las distancias para agrupar los datos
    - **El árbol es una visualización del clustering!**
  - **Parsimonia, Verosimilitud (likelihood)**
    - Evalúa todas las posibilidades para encontrar el árbol más parsimonioso o verosímil (Brute Force, NP-Hard)
    - Permite postular características de las secuencias ancestrales
    - Hay heurísticas

Parsimonia, Verosimilitud

Parsimonia

El mejor árbol es el que explica la historia evolutiva mediante el *menor* número de cambios (es el más parsimonioso)

Verosimilitud

Es una función de los parámetros de un modelo estadístico

La *probabilidad* de un *evento* dado un *conjunto de parámetros*, es la *verosimilitud* del *conjunto de parámetros* dado el *evento*.

Métodos basados en distancias (data clustering)

- **Cómo inferir la filogenia?**
  - **Definir los caracteres a seguir**
  - **Construir una matriz de distancias**
  - **Seleccionar un algoritmo para reconstruir la filogenia a partir de los datos de distancias**

- **Caracteres y estados**
  - **Los caracteres deben evolucionar en forma independiente**
  - **Los estados observados comparten un origen común**

Para secuencias de ADN un caracter corresponde a una posición en la secuencia y los estados posibles, son los nucleótidos A, T, C, G.

Tipos de caracteres

MORFOLÓGICOS Medidas Corporales Medidas Parciales Presencia de estructuras	MOLECULARES Hibridación DNA-DNA RFLP Secuencias (DNA ó Proteínas)
CONTINUOS Medidas Corporales Medidas Parciales Hibridación de DNA-DNA	DISCRETOS Presencia de estructuras RFLP Secuencias (DNA ó Proteínas)

Métodos basados en distancias

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

Matriz de caracteres

Matriz de distancias

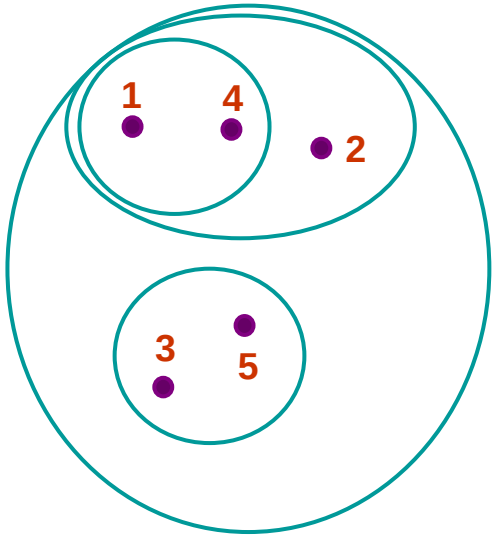
	Sp. 1	Sp. 2	Sp. 3	Sp. 4
Sp. 1	0			
Sp. 2	4	0		
Sp. 3	5	5	0	
Sp. 4	6	4	2	0

Pero hay muchas distancias

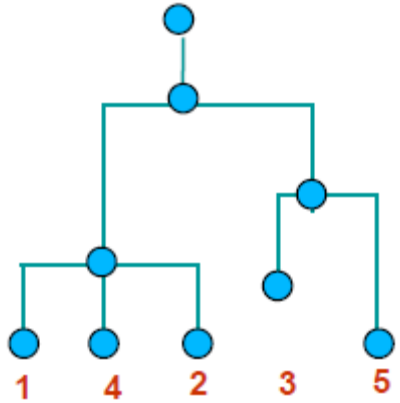
- Fracción de sitios que difieren entre las dos secuencias
- Fracción de sitios que no son similares según BLOSUM62
- Fracción de sitios que no son similares según un Profile específico …

Algoritmos basados en distancias

- Los pares de secuencias más cercanos (neighbors) comparten un ancestro común y están unidos a él por ramas
- El objetivo del método es encontrar un árbol que acomode a todos los *vecinos* correctamente
- El largo de las ramas tiene que concordar con los datos de distancia

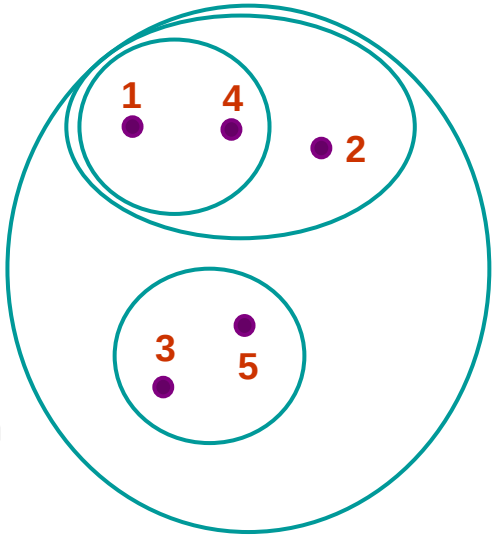


- Usan métodos de *clustering* para agrupar *vecinos*

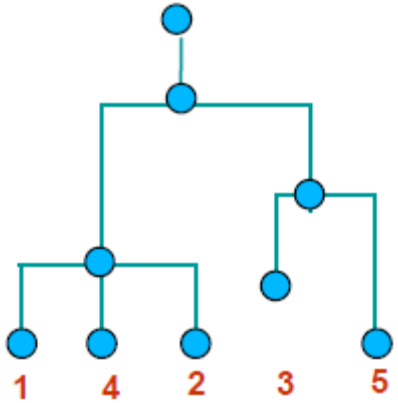


UPGMA

Ultrametric **p**air **g**roup **m**ethod using **a**rithmetic averages (Average linkage)



- **Inicialización**
  - Poner a cada secuencia en su grupo **Gi**
- **Iteración**
  - Encontrar dos grupos **Gi**, **Gj** de manera que la distancia dij sea mínima
  - Definir un nuevo grupo **Gk** (**Gi U Gj**)
  - Definir un nuevo nodo que conecte a **Gi** con **Gj** en el grafo y ponerlo a una altura **= dij/2**
  - Borrar **Gi, Gj**
- **Terminación**
  - Cuando queden sólo dos grupos **i,j**
  - Poner la raíz a una altura **dij/2**



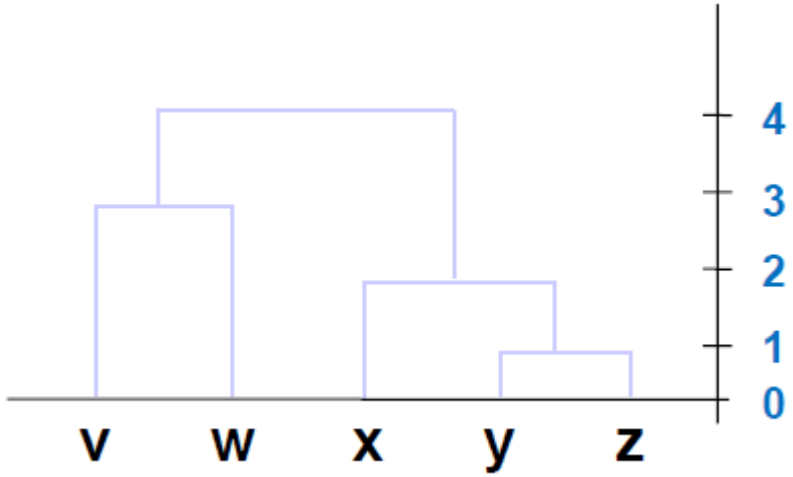
Ejemplo UPGMA

1	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

2	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0

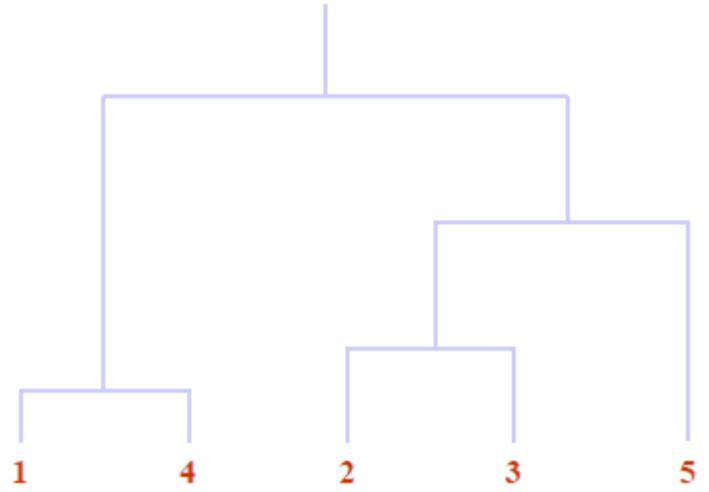
3	v	w	xyz
v	0	6	8
w		0	8
xyz			0

4	vw	xyz
vw	0	8
xyz		0



Distancias ultramétricas

Average Linkage produce distancias ultramétricas



- Distancia métrica

$$d(x,y) > 0 \quad for\ x \neq y$$

$$d(x,y) = 0 \quad for\ x = y$$

$$d(x,y) = d(y,x) \quad \forall\ x,y$$

$$d(x,y) \leq d(x,z) + d(y,z) \quad \forall\ x,y,z \quad \text{(triangle inequality)}$$

- Una distancia es ***ultramétrica*** si además se cumple que

$$d(x,y) \leq max\ (d(x,z),\ d(y,z))$$

- Los árboles ultramétricos se caracterizan por la siguiente propiedad:
  - Tres puntos cualquiera x, y, z pueden ser renombrados de manera que

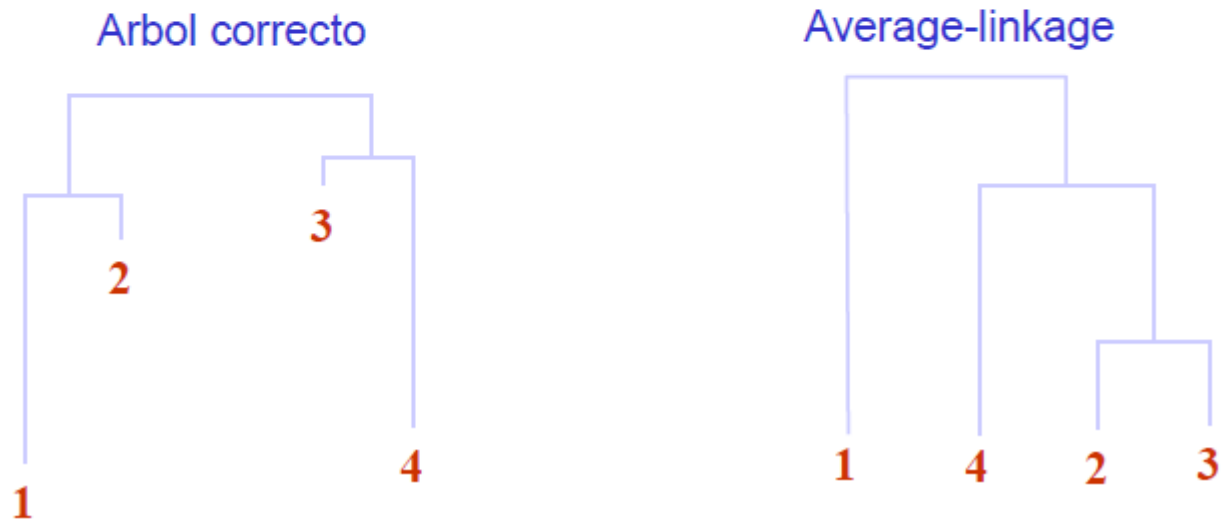
$$d(x,y) \leq d(x,z) = d(y,z)$$

Usar una distancia ultramétrica implica asumir un reloj molecular

La tasa de mutación se asume igual para todas las especies

Problemas con UPGMA

- Cuando la hipótesis del reloj molecular es incorrecta, hay problemas
  - No todas las especies evolucionan a la misma velocidad
  - En estos casos UPGMA (Average linkage) produce resultados incorrectos



Hay otros algoritmos similares

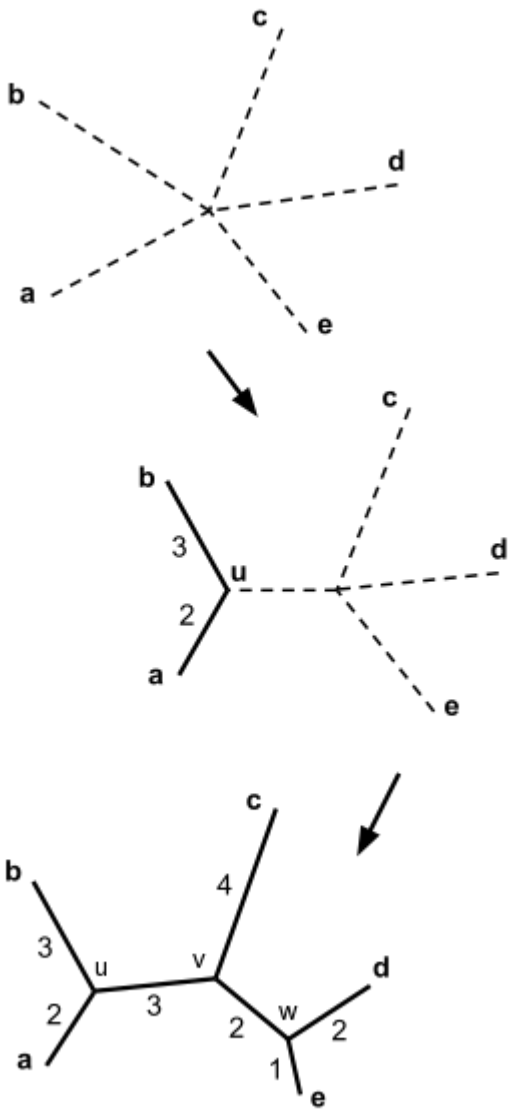
Neighbor joining:

Parecido a UPGMA (usa matriz de distancias), *pero* produce árboles sin raíz.

Inicia con todas las especies en un árbol con topología de estrella.

Transforma la matriz de distancias en cada paso usando la formula:

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

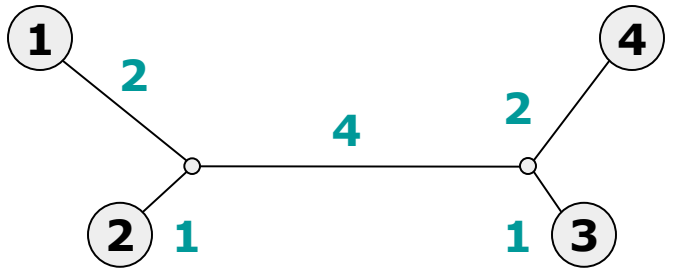


https://en.wikipedia.org/wiki/Neighbor_joining

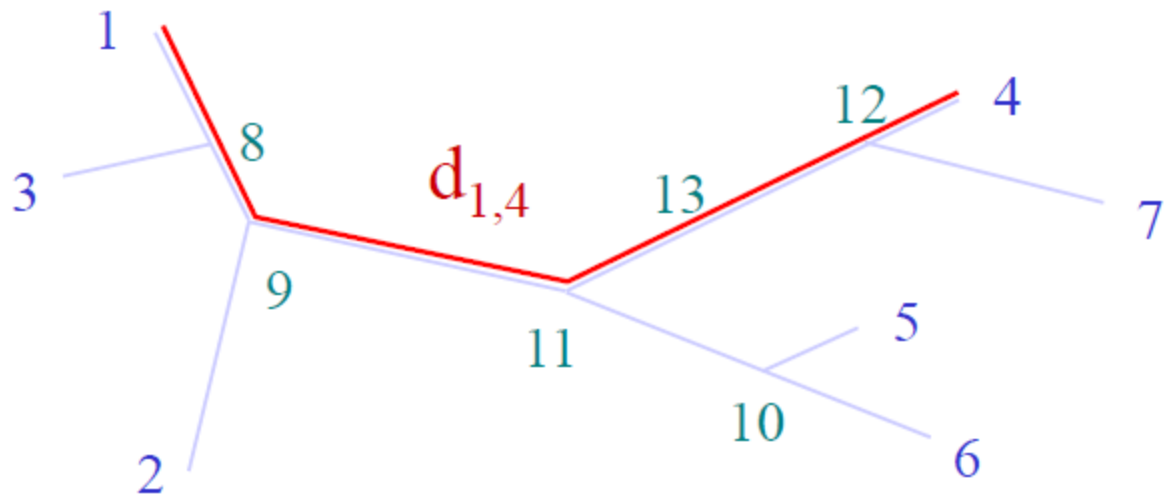
Distintos tipos de distancias

- **Aditivas**
  - **La suma de las longitudes de las ramas de dos especies con su nodo ancestral es igual a la distancia calculada entre las especies**

	Sp. 1	Sp. 2	Sp. 3	Sp. 4
Sp. 1	-			
Sp. 2	3	-		
Sp. 3	7	6	-	
Sp. 4	8	7	3	-



Distancias aditivas



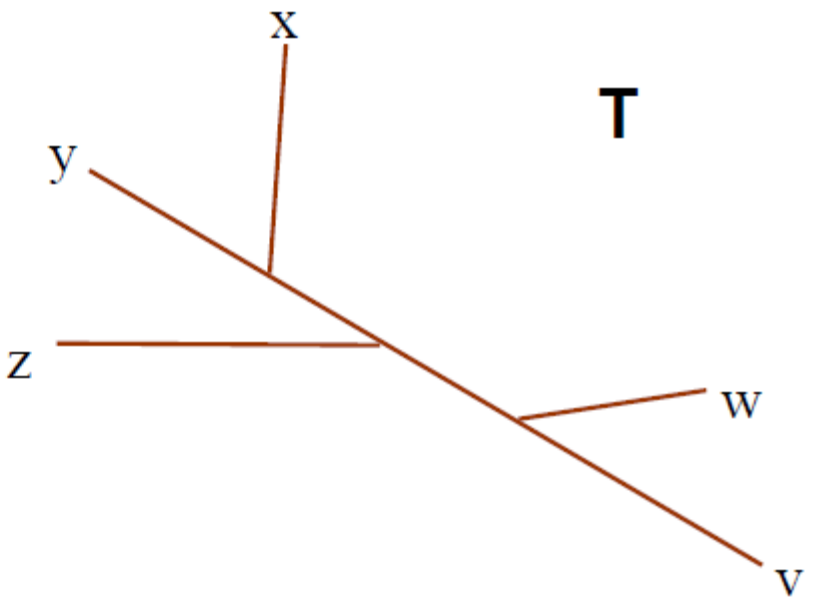
- Es necesario conocer el árbol (la topología)
- Y las distancias entre cada para de *hojas* del árbol
  - La matriz de distancias
- Método
  - Encontrar dos hojas del árbol *i*, *j* con un nodo ancestral *k*
  - Poner el nodo *k* a una distancia de un nodo *m* de manera que
    - $d_{km} = \frac{1}{2} (d_{im} + d_{jm} - d_{ij})$

Additive distances: example

D

	v	w	x	y	z
v	0	10	17	16	16
w		0	15	14	14
x			0	9	15
y				0	14
z					0

T



Si conocemos T y D pero no conocemos las longitudes de cada rama, podemos reconstruirlas utilizando ***additive distances***

Additive distances: example

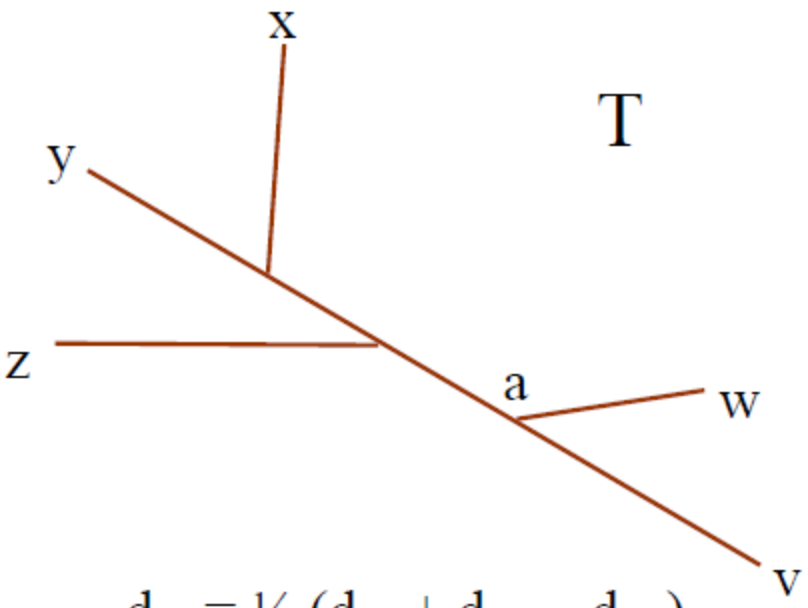
D

	v	w	x	y	z
v	0	10	17	16	16
w		0	15	14	14
x			0	9	15
y				0	14
z					0

D_1

	a	x	y	z
a	0	11	10	10
x		0	9	15
y			0	14
z				0

T



$$d_{ax} = \tfrac{1}{2} (d_{vx} + d_{wx} - d_{vw})$$

$$d_{ay} = \tfrac{1}{2} (d_{vy} + d_{wy} - d_{vw})$$

$$d_{az} = \tfrac{1}{2} (d_{vz} + d_{wz} - d_{vw})$$

Additive distances: example

D_1

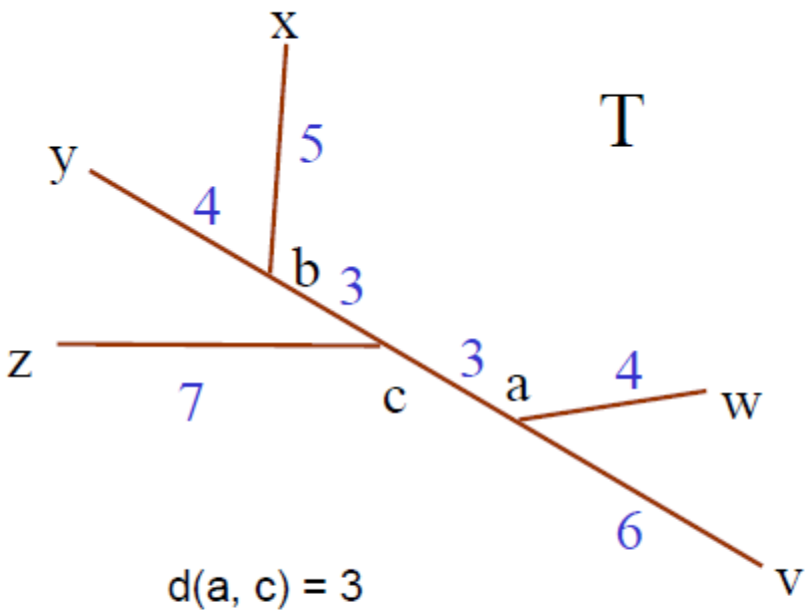
	a	x	y	z
a	0	11	10	10
x		0	9	15
y			0	14
z				0

D_2

	a	b	z
a	0	6	10
b		0	10
z			0

D_3

	a	c
a	0	3
c		0



d(a, c) = 3
d(b, c) = d(a, b) – d(a, c) = 3
d(c, z) = d(a, z) – d(a, c) = 7
d(b, x) = d(a, x) – d(a, b) = 5
d(b, y) = d(a, y) – d(a, b) = 4
d(a, w) = d(z, w) – d(a, z) = 4
d(a, v) = d(z, v) – d(a, z) = 6

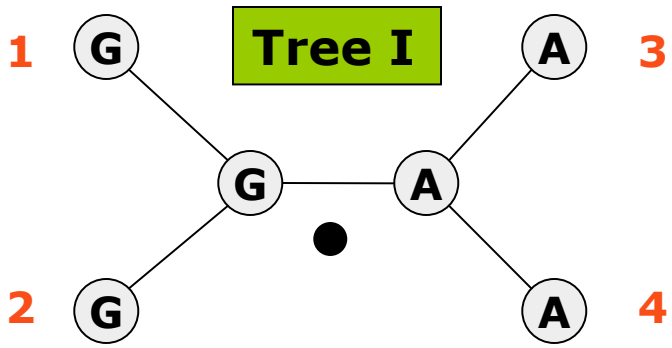
Máxima parsimonia

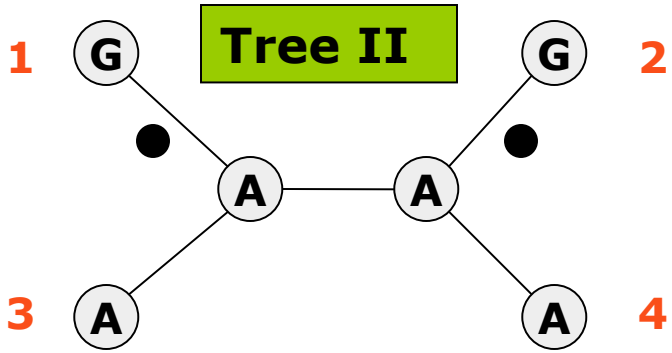
- Predicen el árbol (o árboles) que minimizan el número de cambios (o pasos) que es necesario hacer para generar la variación observada entre las secuencias
- También conocido como *método de evolución mínima*

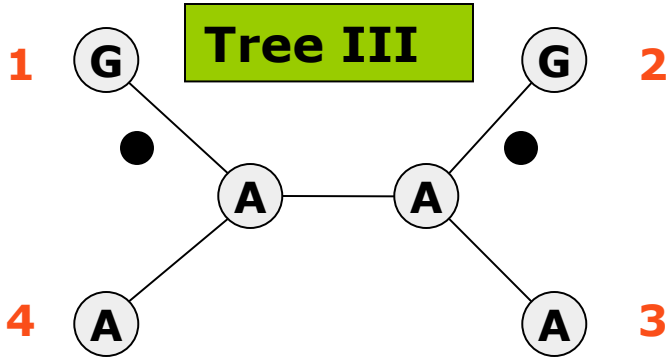
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

- Ejemplo
  - **Para ser informativo un sitio debe tener dos estados presentes en al menos dos especies**
  - **Sitios no informativos: 1, 2, 3, 4, 6 y 8**
  - **Sitios informativos: 5, 7 y 9**
  - **Sólo se analizan los sitios informativos**

Máxima parsimonia: ejemplo







- Hay 3 árboles posibles (sin raíz) para describir la evolución de 4 especies
- Menor número de cambios para explicar la evolución: árbol 1 (1 cambio)
- El mismo análisis se repite para cada uno de los sitios informativos
- El resultado es el árbol que provee el menor número de pasos para acomodar los datos en los sitios informativos (el más parsimonioso)

Máxima parsimonia: detalles

- **Asume que la velocidad de evolución es similar en todas las ramas**
  - **La inferencia obviamente falla cuando esto no se cumple**
  - **Ejemplo: cambio de G a A en forma independiente en dos especies**
    - Especie 1: G > A
    - Especie 2: G > C > T > G > C > A
- **Se pueden asignar puntajes a los árboles**
  - **En lugar de contar cambios se pueden asignar distintos valores a los cambios (por ejemplo usando una matriz)**
- **A diferencia de los métodos de distancia, el método permite obtener la secuencia postulada de cualquier ancestro**

Maxima verosimilitud

- **Maximum likelihood**
  - **Similar al método de máxima parsimonia: usa todas las columnas del alineamiento, considera todos los árboles posibles**
  - **Usa probabilidades**

Heurísticas

Maximum Parsimony/Likelihood methods are computationally hard (brute force)

- **Una forma naïve de identificar el árbol más parsimonioso es por simple enumeración**
  - **Considerar todos los árboles posibles**
  - **Buscar el/los árboles con menor score**

Heurísticas:

- ***Gradient descent (steepest descent)*** **o** ***Hill-climbing***
- **acoplado a**
- ***Tree rearrangement***

Se busca optimizar una función (likelihood) mediante exploración limitada del espacio de árboles posibles.

Hill Climbing / Gradient Descent

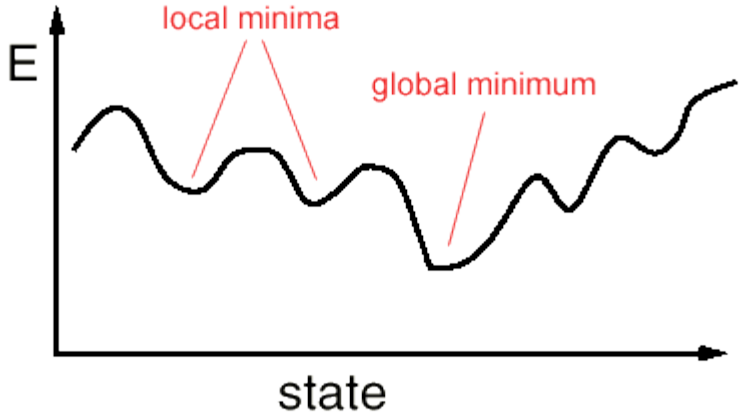
Algoritmo iterativo

Inicialización:

Comienza con una solución ***arbitraria*** al problema

Iteración:

Se buscan mejores soluciones a partir de mejorar una métrica o score, cambiando una parte de la solución.



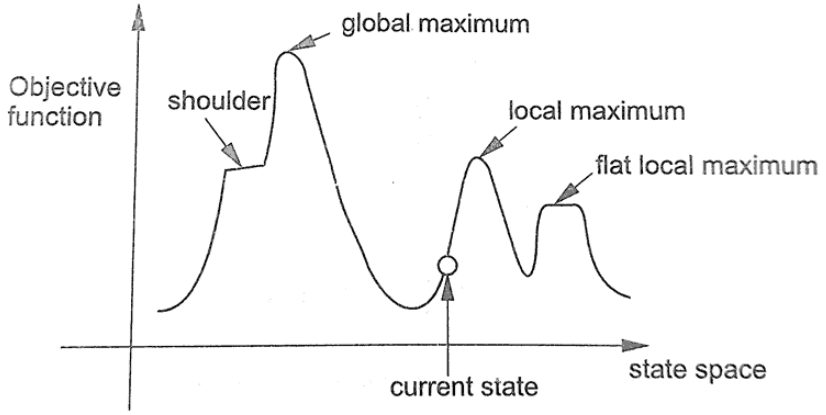


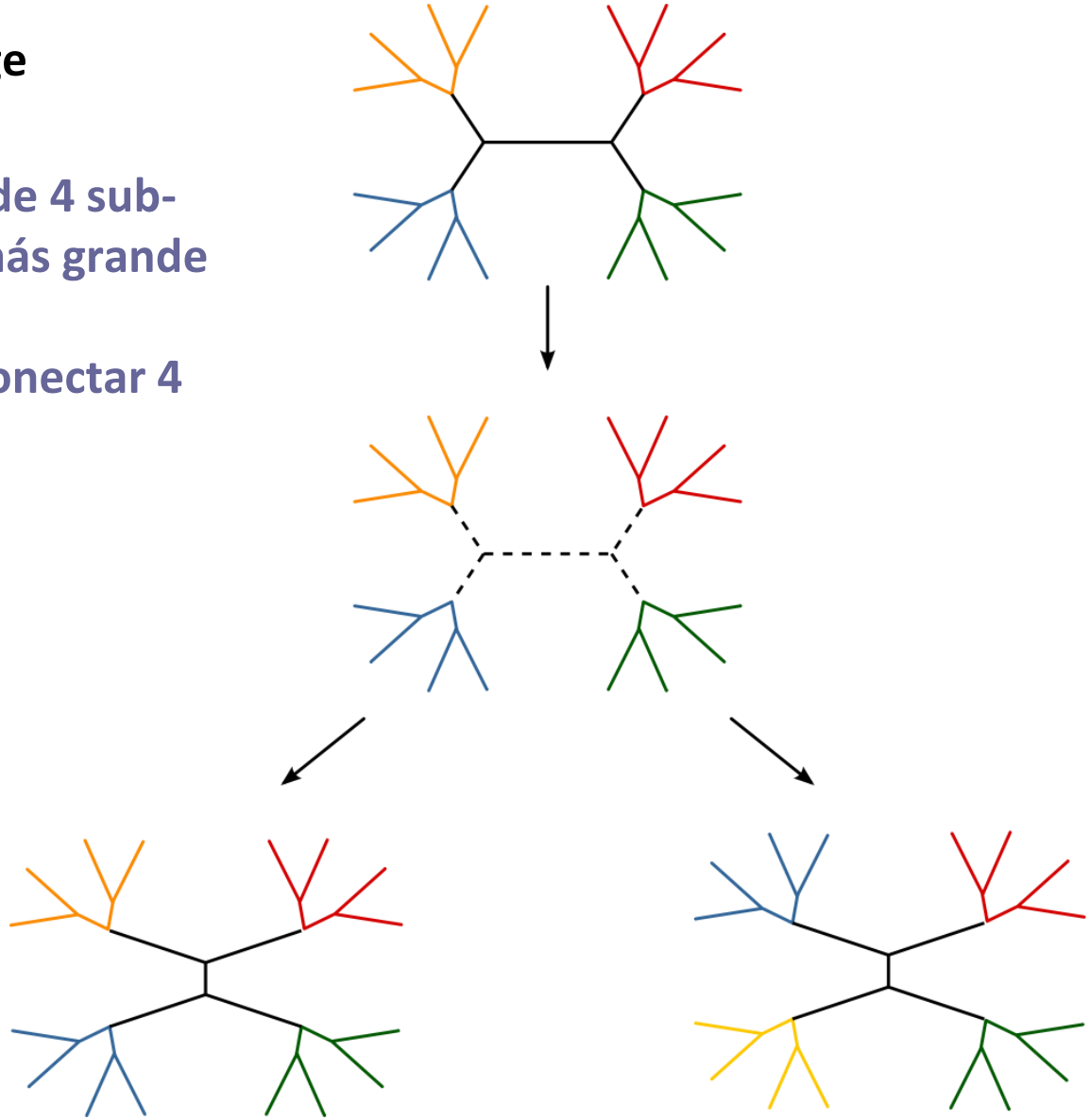
Fig: Hill Climbing

Tree rearrangements

Nearest Neighbor Interchange

Intercambia la conectividad de 4 sub-árboles dentro de un árbol más grande

Hay 3 maneras posibles de conectar 4 objetos.

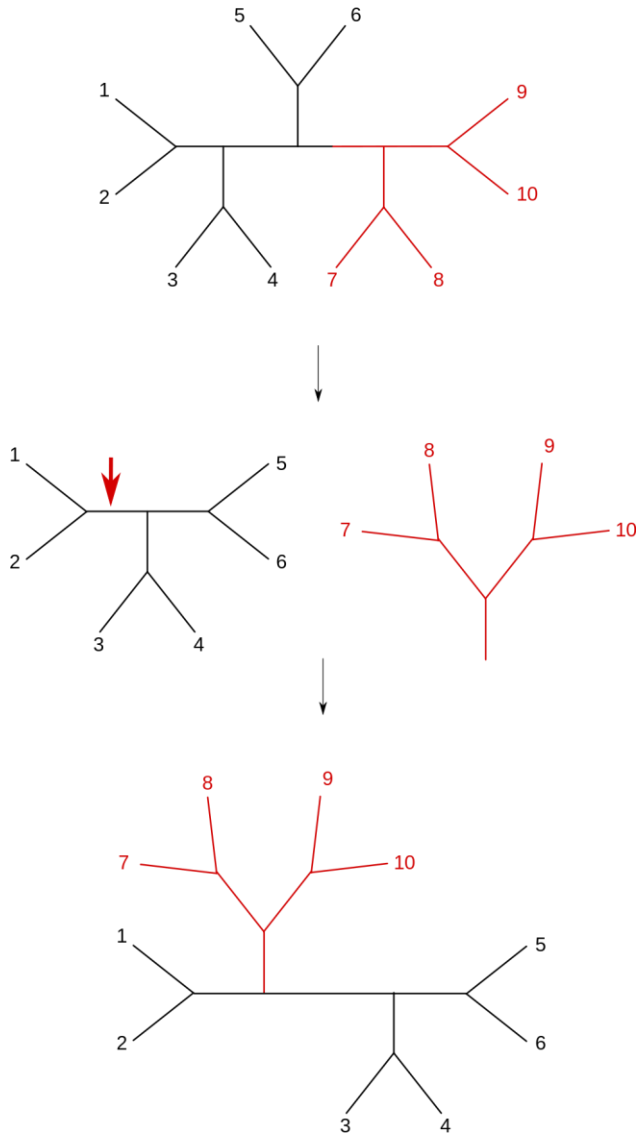


Tree rearrangements

Subtree pruning and regrafting

(Podado y re-enraizado de subárboles)

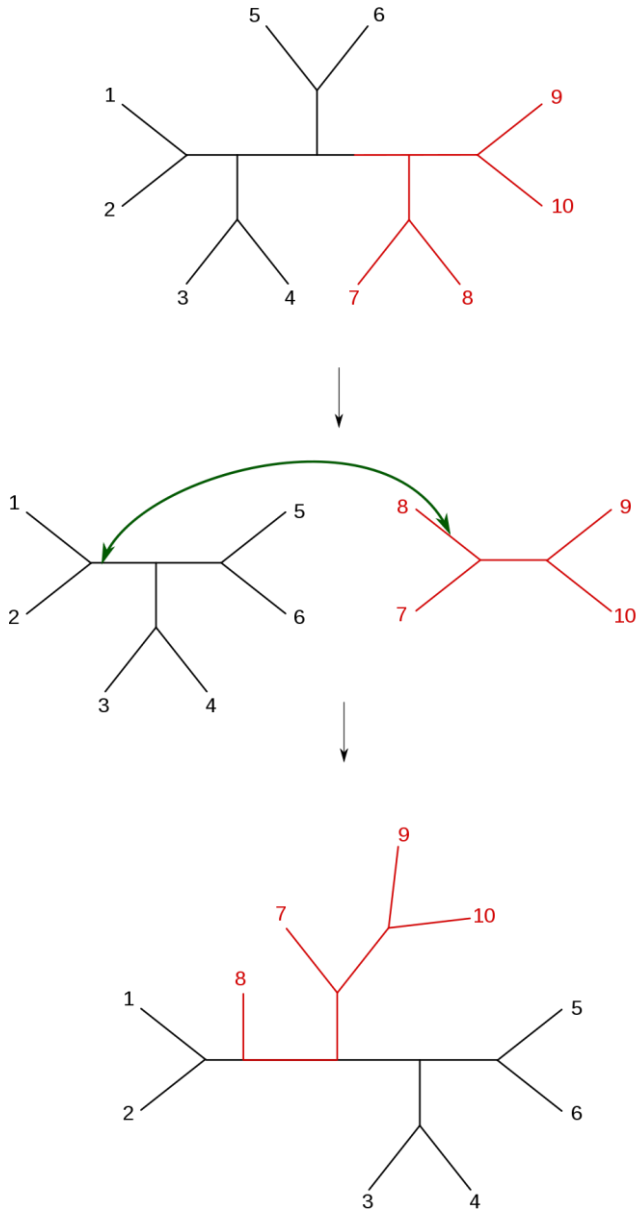
Selecciona un sub-árbol, que se remueve del árbol original y se pone en cualquier otro lugar del árbol.



Tree rearrangements

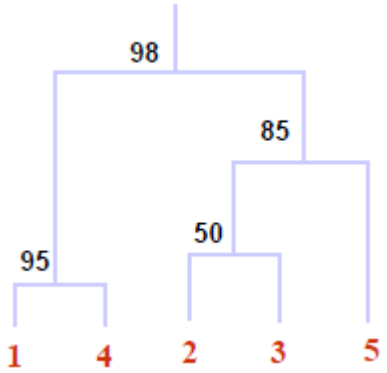
Tree bisection and reconnection

Se remueve un sub-árbol de un nodo interior del árbol original y se intentan todas las posibles conexiones entre los dos árboles que se generan



Testeo de árboles

- **Bootstrap test**
  - **Bootstrap resampling technique (Efron 1982)**
  - **Dado un número de secuencias M de longitud N (un alineamiento), y un árbol calculado por un método cualquiera, se genera un nuevo set de secuencias M' en el cual N' bases/residuos elegidos al azar son reemplazados, también al azar.**
  - **En base a este nuevo set M' se recalcula el árbol utilizando el mismo método y se comparan las topologías del árbol.**
  - **Esto se repite varias veces (100, 1000 repeticiones) y se calcula, para cada rama un valor de bootstrap**
  - **Bootstrap value: % de veces que la rama aparece en los distintos árboles**
  - **Bootstrap values >= 95% corresponden a ramas “correctas”**



Testeo de árboles: Bootstrap

El test de bootstrap original (Efron, 1982; Felsenstein 1985) funciona bien para ***pocos datos***.

Para datasets más grandes (cientos de miles de taxones o secuencias), los valores de bootstrap tienden a ser bajos.

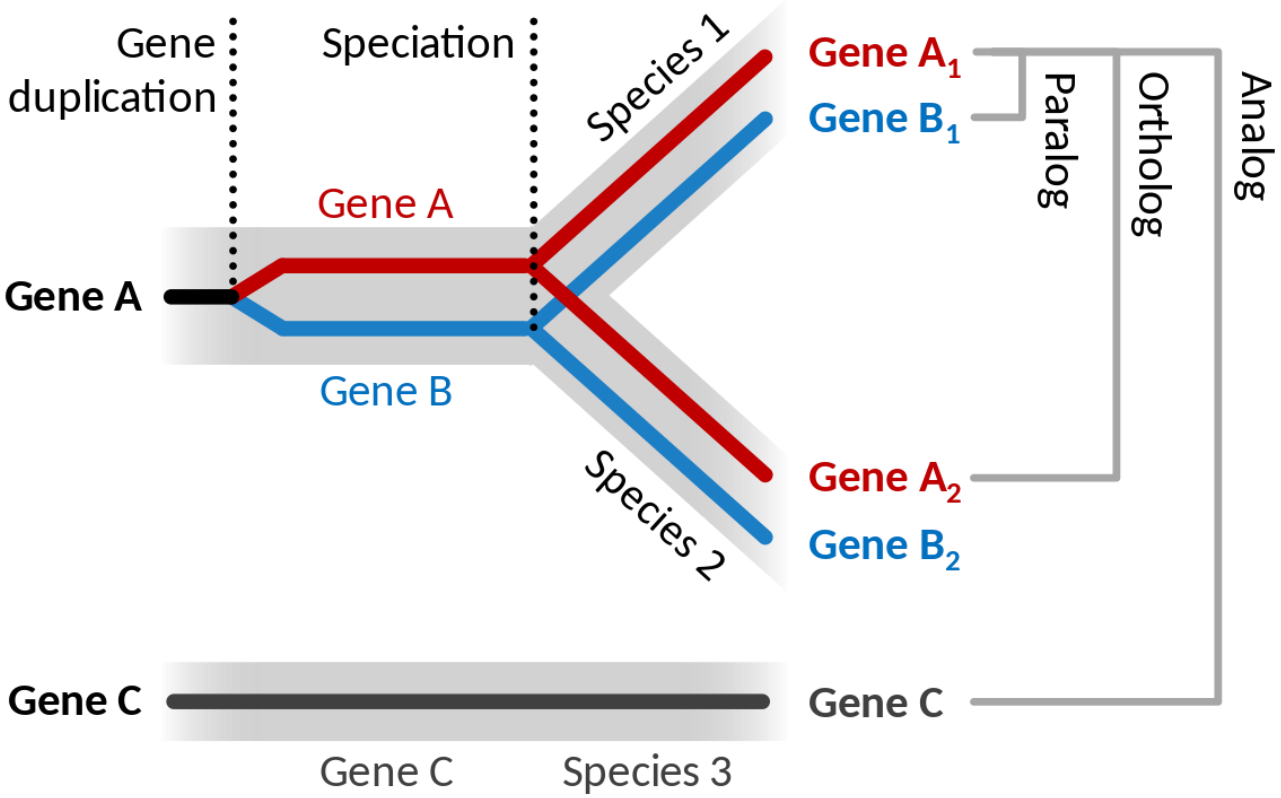
TBE – Transfer Bootstrap Expectation, es una variante que permite calcular valores de soporte usando la técnica de bootstrap, pero donde cambia el concepto de presencia/ausencia de ramas en las replicas y permutaciones.

Lemoine F, Domelevo Entfellner JB, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature. 2018 Apr;556(7702):452-456. doi: 10.1038/s41586-018-0043-0. Epub 2018 Apr 18. PMID: 29670290; PMCID: PMC6030568.

Testeo de árboles

- **Jacknife**
  - **Muy similar al test de bootstrapping**
  - **Se generan nuevos data sets por muestreo parcial del original**
  - **Usualmente se muestrea el 50% de los datos originales**
  - **Se rehacen los árboles y se verifica la topología**
  - **Se hacen varios re-muestreos (100-1000 veces)**
  - **Se construye un árbol consenso con valores de confidencia para cada rama**

Ortólogos y parálogos

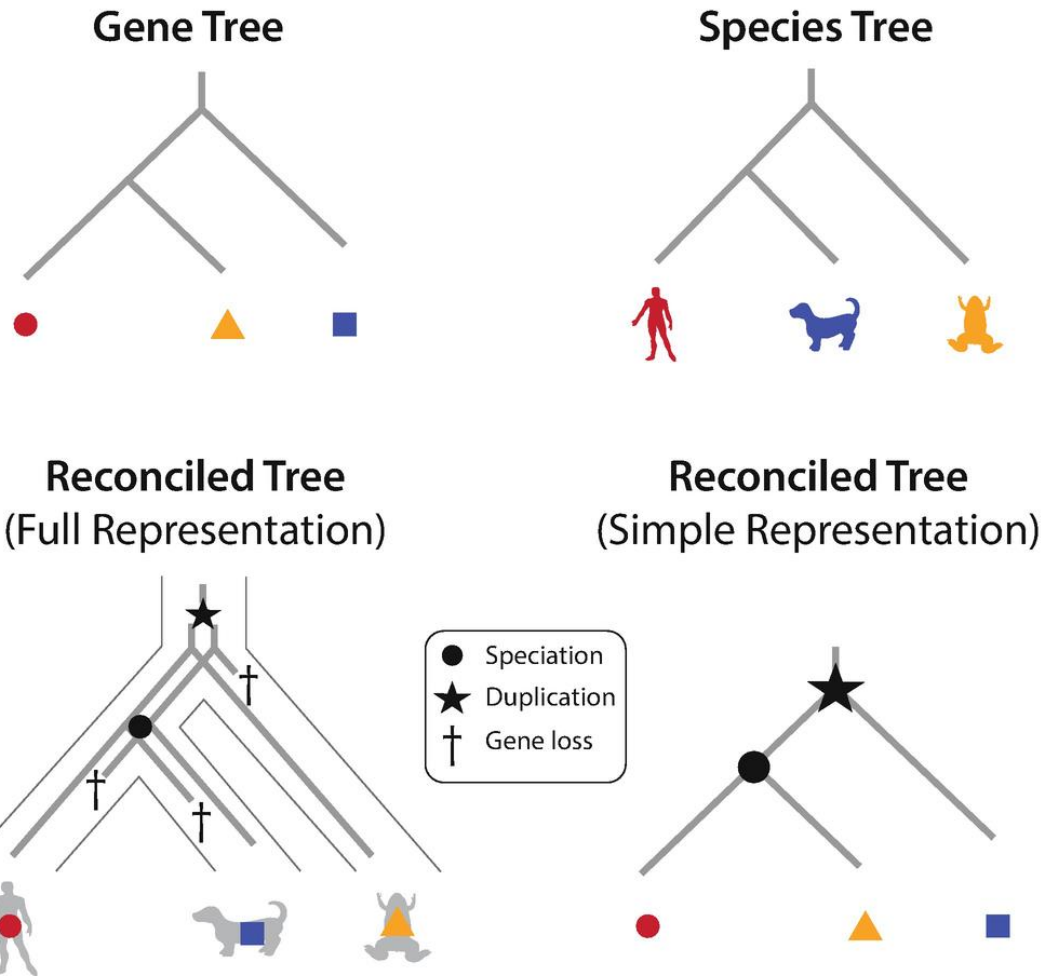


Parálogos: homologos relacionados por eventos de duplicación génica desde el ultimo ancestro común

Ortólogos: homologos relacionados por eventos de especiación

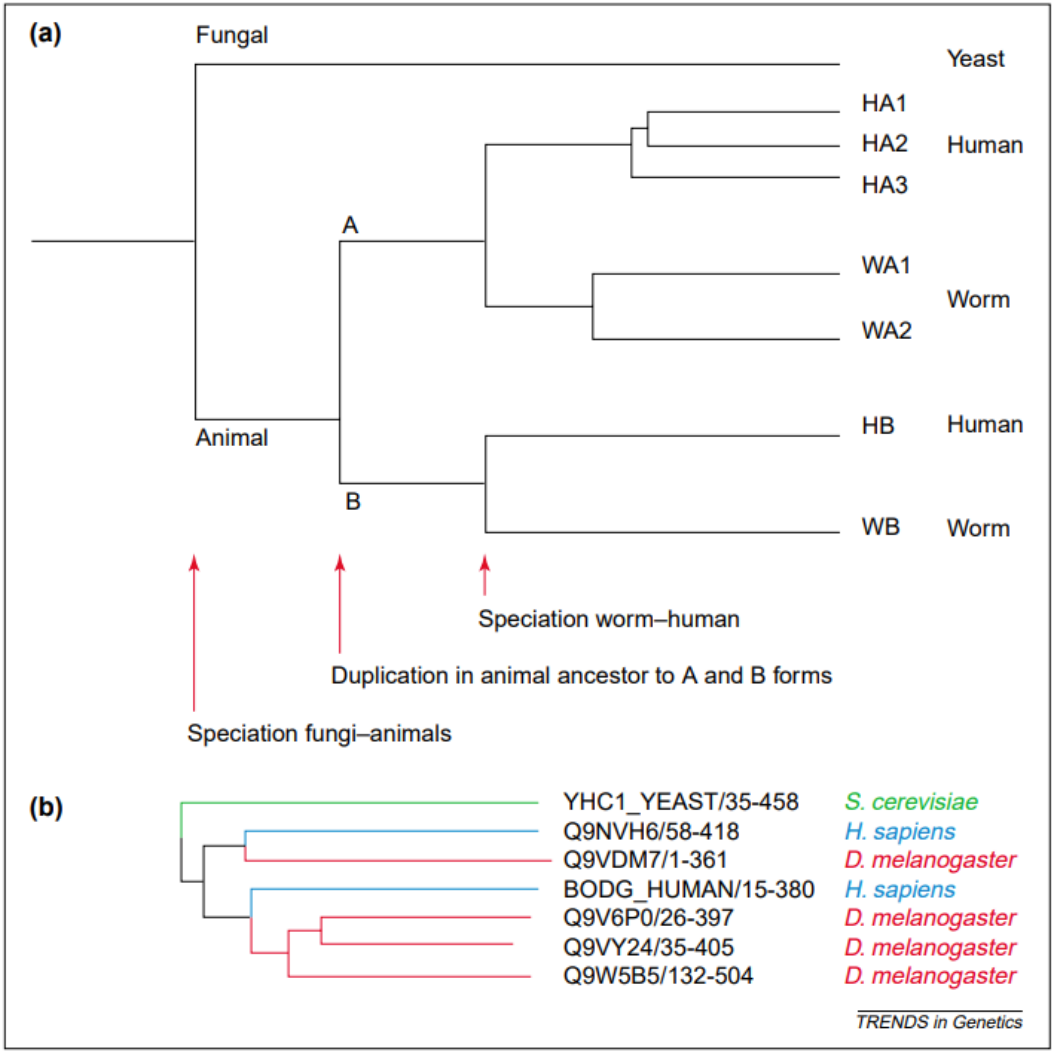
https://en.wikipedia.org/wiki/Sequence_homology

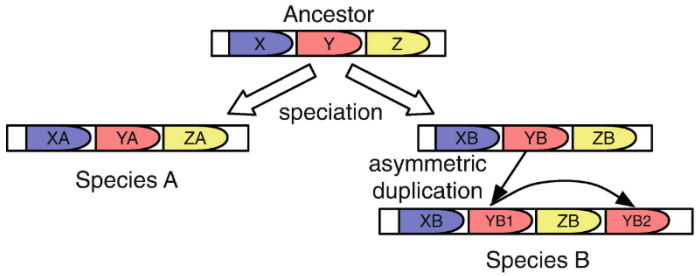
Aplicaciones de los árboles filogenéticos



Los árboles filogenéticos nos ayudan a inferir escenarios e hipótesis evolutivas

Ortologos y Parálogos

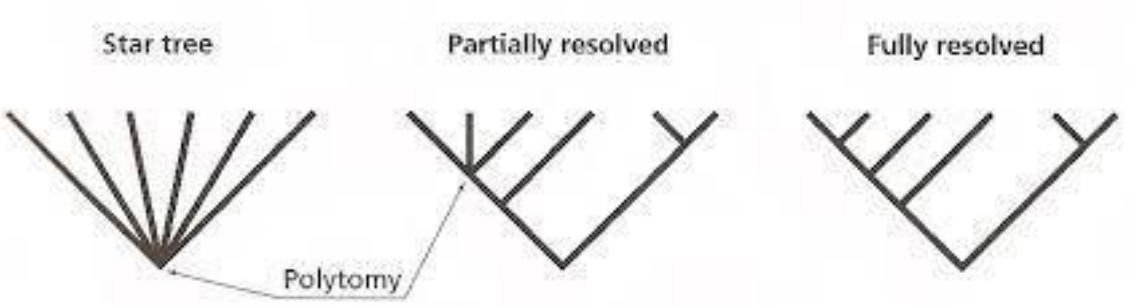




Inparalogs vs Outparalogs

Ramas no resueltas: politomías

una **politomía** es una sección de un árbol filogenético en la que las relaciones no se pueden resolver totalmente en una serie de dos vías o divisiones (dicotomía).



Paquetes de software

- **Phylip**
  - **Unix, linea de comando. Gratuito.**
  - **DNA, Proteinas,**
  - **Distancias, Parsimonia**
  - **Bootstrap, Jacknife**
- **PAUP**
  - **Similar a Phylip. Comercial. Interfase gráfica, linea de comando.**
- **PhyML**
  - **Maximum likelihood**
- **RAxML / ExaML, RAxML-NG**
  - **RAxML, maximum likelihood (ML) tree inference**
  - **ExaML, genome-scale datasets**
  - **RAxML-NG, Greedy tree search algorithm**