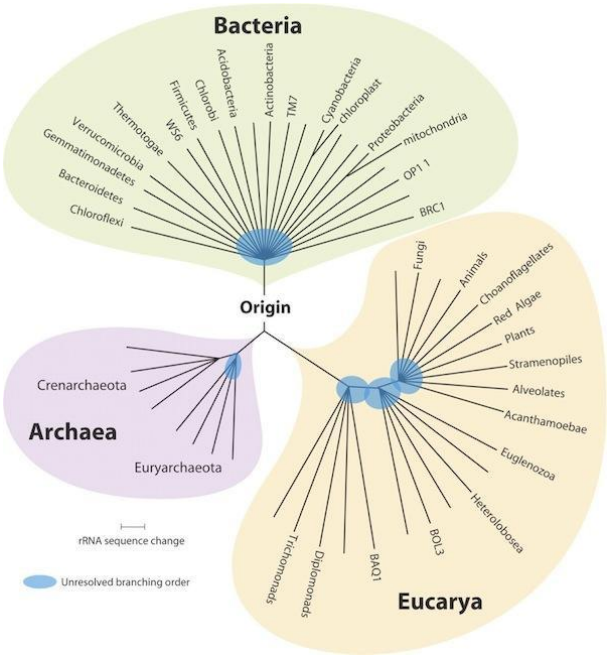
Curso de Evolución 2023
Facultad de Ciencias
Montevideo, Uruguay
http://eva.fcien.udelar.edu.uy/
https://www.youtube.com/@CursoEvolucion/videos



Tema 2. Las filogenias como contexto de análisis de la evolución. Aplicaciones del análisis filogenético. Análisis filogenético según el principio de parsimonia. Métodos basados en distancias y en modelos de evolución molecular.

Algunos métodos de inferencia filogenética

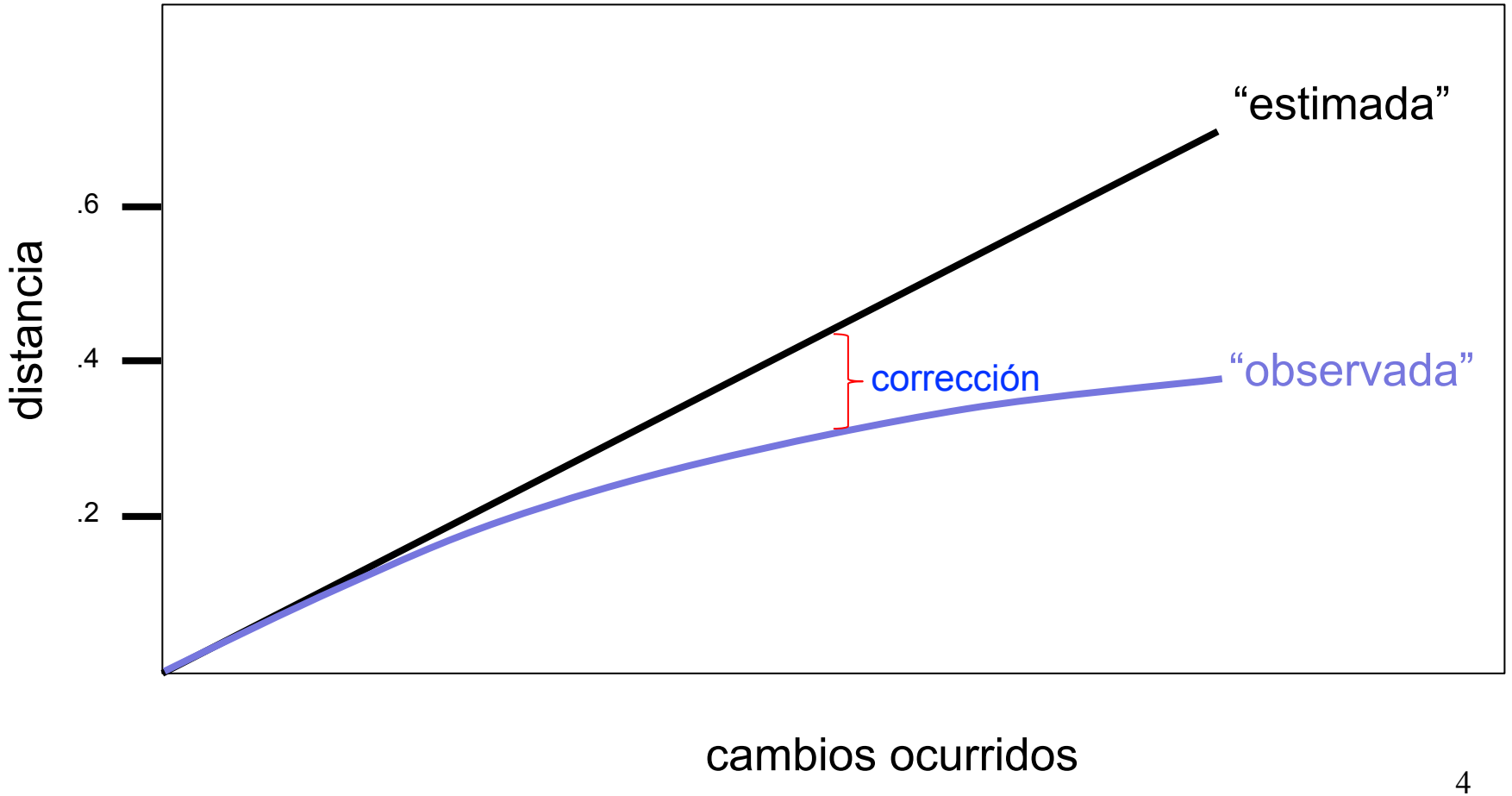
Método	Variantes	Criterio de optimización	Uso de variación no observada
Parsimonia	máxima parsimonia	minimizar el número de pasos requeridos para obtener los datos	No
Distancias	- evolución mínima - unión de vecinos (neighbor joining)	minimizar los cambios requeridos para obtener las distancias estimadas entre taxa una aproximación al árbol de evolución mínima	Sí (incorporadas en las distancias) ídem
Inferencia estadística	…		

Optimización: parsimonia vs. distancias

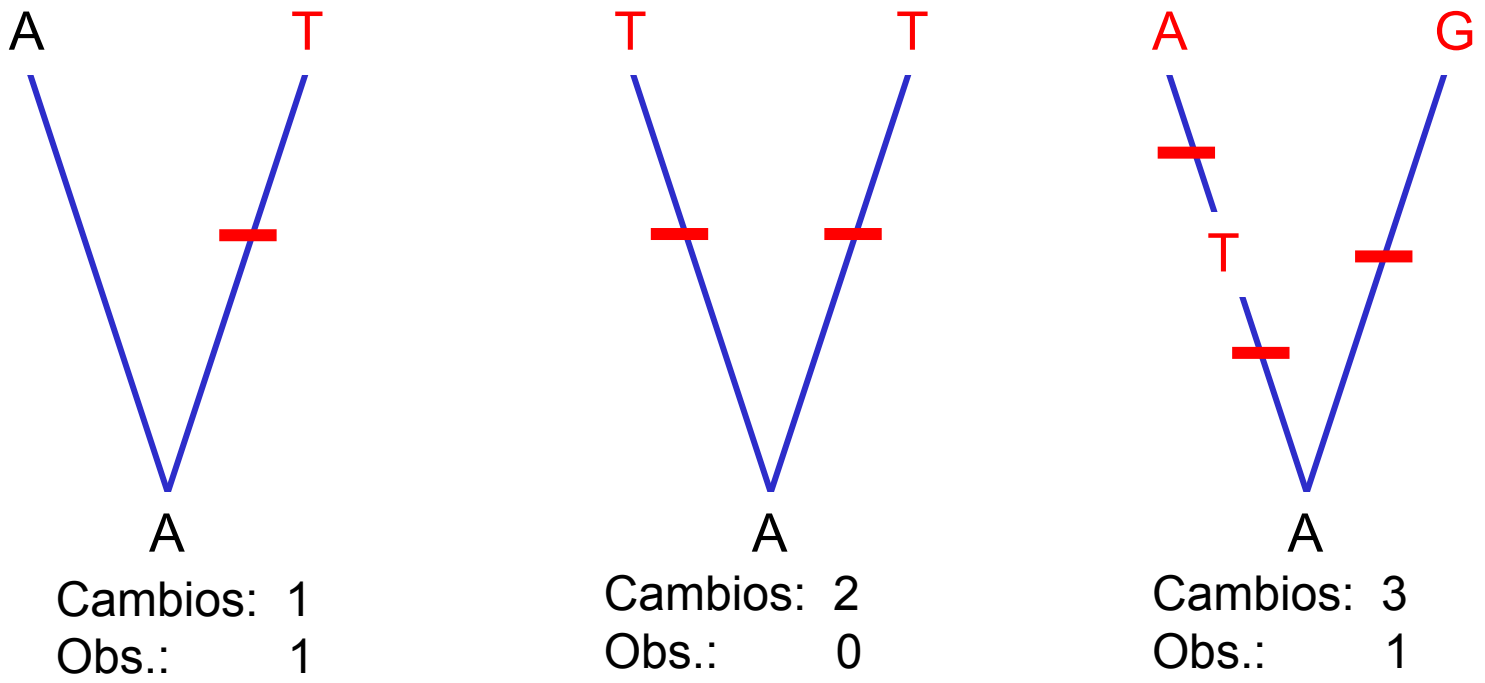
	Parsmionia	Distancias
Criterio de optimización	Minimizar la longitud (número de pasos) del árbol	Minimizar la longitud (suma de todas las ramas, medidas como distancias) del árbol
Efecto de la homoplasia*	Los mejores árboles requieren más pasos que el mínimo ideal	1) Las distancias estimadas son mayores a las observadas.
		2) El árbol óptimo requiere un largo total mayor al mínimo requerido por las distancias.

Homoplasia: similitud no heredada de un ancestro común.
En contraste con homología: similitud resultante de una condición heredada de un ancestro común.
Estos conceptos se aplican a caracteres, así como a estados de caracteres.

¿Por qué calcular distancias?



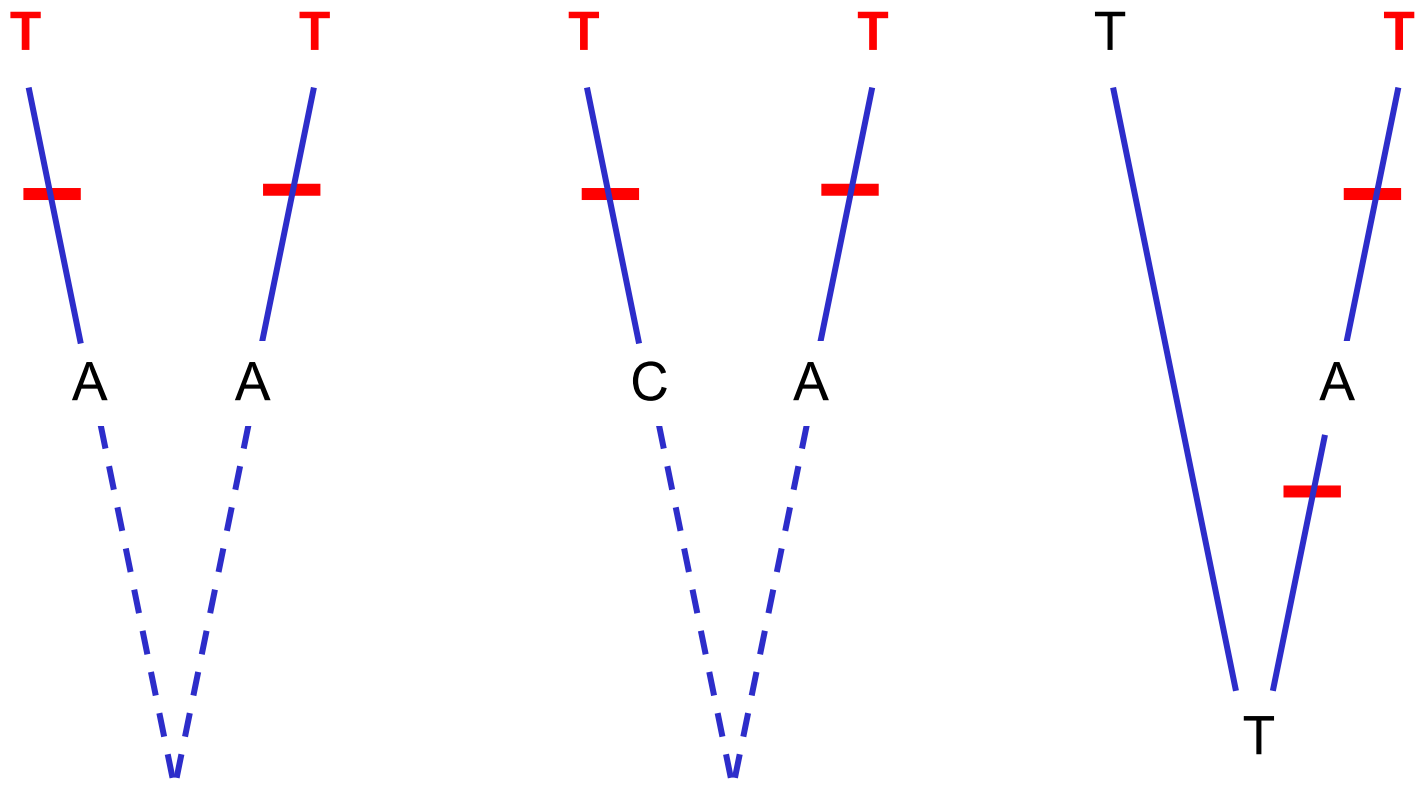
Distancia observada ≤ “Distancia real”



“distancia real”: cambios acumulados en la evolución

Distancia observada: diferencias observadas entre las secuencias finales

Homoplasia:
paralelismo, convergencia, reversión



Administrando la homoplasia: parsimonia vs. distancias

- Parsimonia:
  - No intenta evaluar (o corregir) cambios no observados como diferencias en los datos.
  - La homoplasia resulta en cambios (pasos) adicionales, o sea árboles más largos al mínimo posible.
  - El criterio de máxima parsimonia procura minimizar la longitud del árbol.
- Distancias:
  - Las distancias corregidas (estimadas) procuran, precisamente, ajustar los cambios no observados,que resultan en homoplasia.

Relación entre distancias y modelos de evolución molecular

- Para calcular la distancia “estimada” (corregida con una estimación de los cambios no observados), necesitamos la distancia observada y un modelo de evolución molecular.

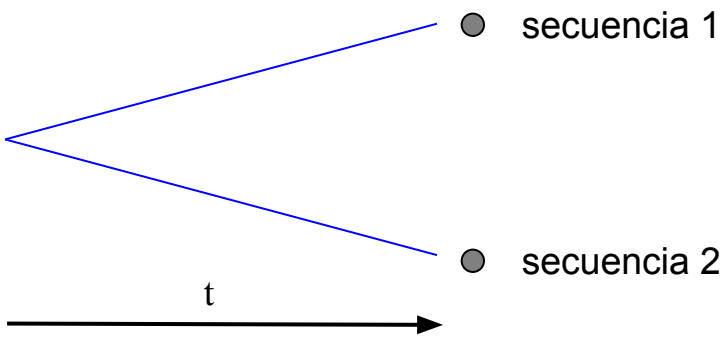
- Ejemplos: modelo de Jukes y Cantor:

	G	A	T	C
G	1-3α	α	α	α
A	α		α	α
T	α	α	1-3α	α
C	α	α	α	1-3α

- La distancia correspondiente es $d = \frac{3}{4} \ln(1 - (4/3)\, p)$, siendo p la distancia observada (proporción de sitios diferentes).

Distancia de Jukes-Cantor

	G	A	T	C
G	1-3α	α	α	α
A	α		α	α
T	α	α	1-3α	α
C	α	α	α	1-3α



Tasa de sustitución $r = 3\alpha$

$E(d) = 2tr$

Por otra parte, t los sitios idénticos entre las dos secuencias (q) en un tiempo se mantienen idénticos en t+1 con probabilidad $(1-r)^2$

- los sitios distintos en t se vuelven idénticos con probabilidad $(\frac{2}{3})r(1-r)$

Como resultado:

$$q_{t+1} = (1 - 2r)q_t + \frac{2}{3}r(1 - q_t)$$

El cambio neto es:

$$q_{t+1} - q_t = \frac{2r}{3} - \frac{8r}{3}q_t$$

Pasando a tiempo continuo, tenemos (ecuación 1):

$$\frac{d_q}{d_t} = \frac{2r}{3} - \frac{8r}{3}q$$

Sabemos que $q_0 = 1$ en t_0, y la solución para esta ecuación diferencial es:

$$q = 1 - \frac{3}{4}(1 - e^{-8rt/3})$$

Esta fórmula nos permitiría estimar q si conociéramos dos variables (r y t) que, en general, nos son desconocidas.

Por otra parte, sabemos que la divergencia (la distancia d que estamos intentando estimar) entre dos secuencias:

$$E(d_{JK}) = 2rt$$

Esta es una de las ecuaciones, tan sencilla como importante, de la evolución molecular. La divergencia es función del tiempo y de la tasa de evolución r de cada una de las ramas que van desde el ancestro común hasta las secuencias que estamos comparando.

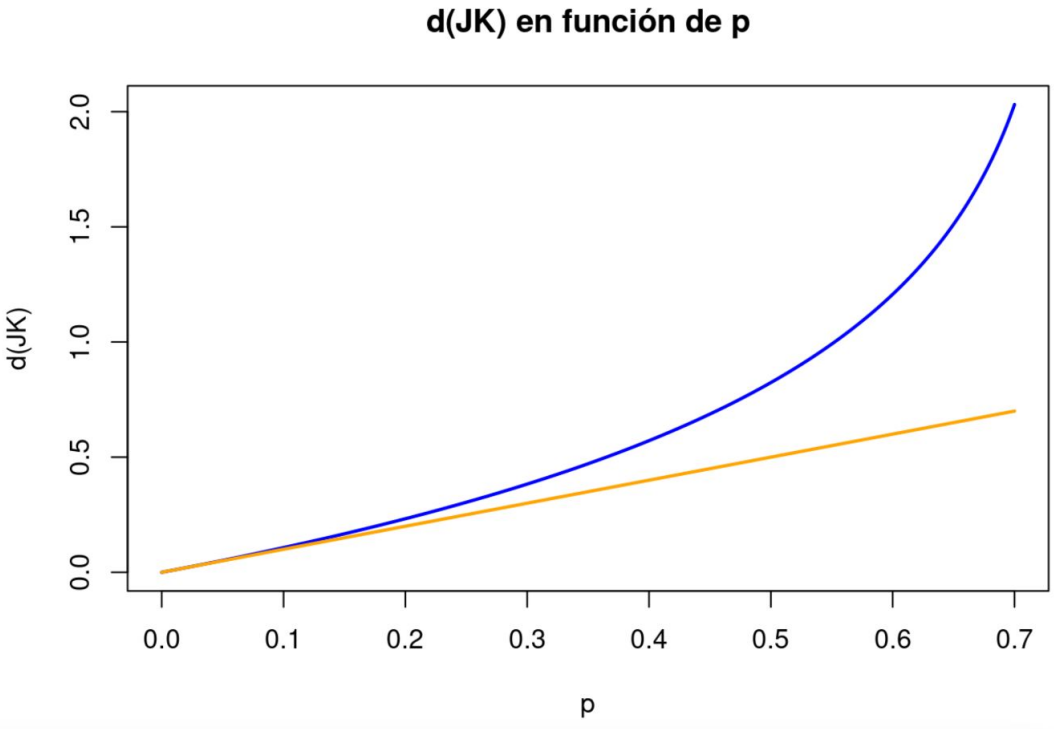
Sustituyendo $2rt$ por d en la ecuación 1 y despejando d obtenemos:

$$E(d_{JK}) = -\frac{3}{4}ln(1 - \frac{4}{3}p)$$

Sustituyendo p (la verdadera divergencia entre las dos secuencias) por $\hat{p}$, nuestra estimación empírica, podemos estimar la distancia de Jukes-Cantor para un par cualquiera de secuencias.

Divergencia observada (p) y distancia JK

p	d
0.010	0.010
0.050	0.052
0.100	0.107
0.200	0.233
0.300	0.383
0.400	0.572
0.500	0.824
0.600	1.207
0.700	2.031



Comentarios sobre distancias

- por debajo de ≈10% de diferencia observada no vale la pena el ajuste

- por encima de ≈ 50% la divergencia es demasiado alta como para estimar apropiadamente una distancia (esperamos que dos secuencias completamente al azar difieran en 75% de los sitios)

Relación entre distancias y modelos de evolución molecular

El modelo de Jukes y Cantor tiene un solo parámetro (μ).

	G	A	T	C
G	1-3α	α	α	α
A	α		α	α
T	α	α	1-3α	α
C	α	α	α	1-3α

Un modelo popular de dos parámetros es el de Kimura, que distingue transiciones y transversiones:

	G	A	T	C
G	1-α -2β	α	β	β
A	α	1-α-2β	β	β
T	β	β	1-α 2β	α
C	β	β	α	1-α-2β

Motivación del modelo de Kimura 2 parámetros

ADN mitocondrial (sitios divergentes)

```
Humano       G G T C T C T A
Chimpancé    A A C T C T C T
```

A, G: purinas

C, T: pirimidinas

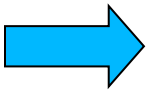
7/8 diferencias son transiciones

1/8 diferencias es una transversión

Filogenias usando distancias

	Caracteres									
	1	2	3	4	5	6	7	8	9	10
Especie A	c	a	a	g	t	c	c	g	t	a
Especie B	.	.	t	.	.	t	.	a	.	.
Especie C	.	.	t	.	.	.	t	a	.	.
Especie D	t	g	.	.	c	.	.	.	.	g
Especie E	t	g	.	a	c	.	.	t	.	.

	A	B	C	D	E
A		0,3	0,3	0,4	0,5
B			0,2	0,6	0,7
C				0,7	0,7
D					0,3



Árbol

Método de unión de vecinos (NJ:“neighbor joining”)

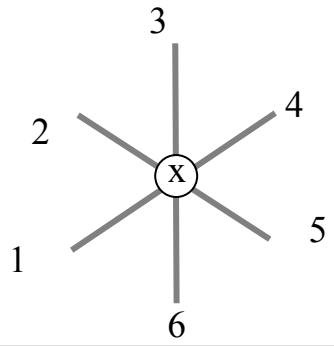
- NJ es una aproximación eficiente al criterio de evolución mínima:

El árbol óptimo es aquel que requiere la menor longitud total (suma de todas las ramas, medidas como distancias moleculares) para representar las distancias estimadas.

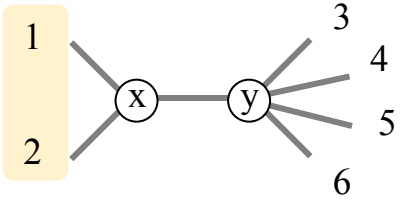
- El algoritmo de NJ es sumamente eficiente: se puede aplicar a gran cantidad de OTUs.
- Al optar por optimizar sobre distancias,
  - Resumimos la información sobre las OTUs (caracteres y sus estados) en distancias entre pares de OTUs. (A diferencia de la parsimonia, que trabaja con los caracteres originales)
  - Las distancias pueden ser las observadas (número o fracción de sitios diferentes) o, más comúnmente, distancias corregidas en base a un modelo de evolución para incorporar cambios no observados. (A diferencia de la parsimonia, que no usa caracteres no observados, y solamente incorpora cambios adicionales cuando así lo requieren los árboles más parsimoniosos).

Método de unión de vecinos (NJ)

Paso 1: Comenzamos con una politomía (todos los taxa unidos a un único nodo interno).

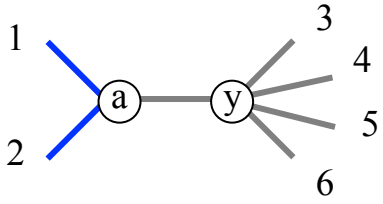


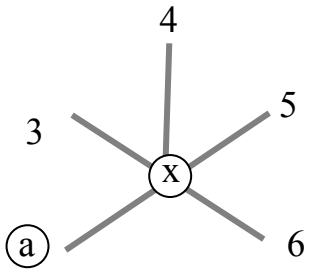
Paso 2: Elegimos un par de "vecinos" a unir: aquel que resulte en el árbol más corto.



(1,2), (3,4,5,6)
(1,3), (2,4,5,6)
….
(5,6) (1,2,3,4)

Paso 3: Reemplazamos dicho par por el nodo que los une (a). Para eso, volvemos a calcular las distancias entre todos los taxones (a y 3, 4, 5, 6 en el ejemplo).

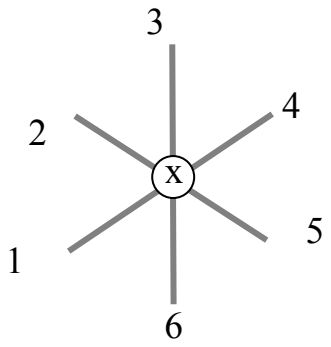




- Vamos al Paso 1 con el árbol obtenido en el Paso 3.
- Repetimos el ciclo (1,2 3) hasta obtener un árbol completamente resuelto.
- Calculamos las longitudes de todas las ramas de dicho árbol.

Método NJ: algunos detalles

- Punto de partida: la hipótesis nula (H_0) para un ejemplo con 6 OTUs (*m* = 6)
- Longitud S_0 (en distancias) del árbol H_0 :



$$S_0 = \sum_{i=1}^{m} L_{iX} = \frac{1}{m-1} \sum_{i<j}^{m} d_{ij}$$

Observamos que

$$d_{12} = L_{1X} + L_{2X}$$

En la suma basada en d_{ij}, L_{1X} forma parte de $d_{12.}$ d_{13} y d_{14}, para un total de *m*-1 veces. Esto se aplica a todos los segmentos L_{iX}, por lo que dividimos la suma por *m*-1.

Método NJ ilustrado con un ejemplo

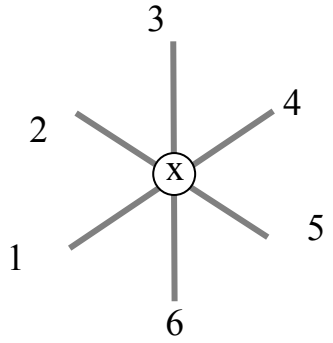
(tomado de Nei & Kumar, 2000. Molecular Evolution and Phylogenetics, Oxford Univ. Press)

$$S_0 = \sum_{i=1}^{m} L_{iX} = \frac{1}{m-1} \sum_{i<j}^{m} d_{ij}$$

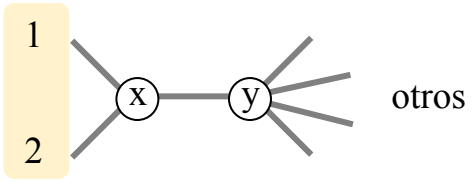
matriz de distancias

	1	**2**	**3**	**4**	**5**
2	9				
3	12	7			
4	15	10	5		
5	20	15	10	11	
6	16	11	6	7	8

S_0 = ⅕ (9 + 12 + …+ 8) = 32.4



- ¿Por qué no usar $S_0 = d_{12} + d_{34} + d_{56}$?
- Aquí y en otras fases del método, usamos toda la información de la matriz en lugar de un subconjunto porque las distancias son estimaciones sujetas a error



Elección del primer par de vecinos

Enumeramos todos los pares candidatos a ser los primeros dos "vecinos" $(1, 2), (1, 3), \ldots. (m-1, m)$ Tomamos (1,2) como ejemplo. El árbol resultante tiene un par terminal $(1, 2)$, una politomía $(3, 4, 5, 6)$, y una rama central (x, y) que une las dos partes anteriores. La longitud total de este árbol es:

$$S_{12} = L_{1x} + L_{2x} + L_{xy} + \sum_{3 \le i < j} L_{ij}$$

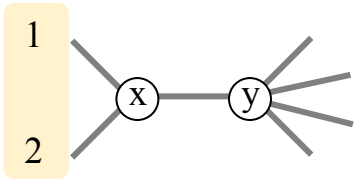
Notamos que $L_{1x} + L_{2x} = d_{12}$

Y tenemos una fórmula para la longitud totad de un árbol politómico (en este caso de 4 taxones, aunque numerados de 3 a 6).

$$S_0 = \frac{1}{m-3} \sum_{3 \le i < j} d_{ij}$$

Nos queda estimar L_{xy}. Lo haremos a partir de la suma de todas las distancias que pasan por xy. Dicha suma pasa 8 veces por xy, e incluye 4 veces $(m-2)$ la distancia d_{12}. Incluye además 2 veces cada rama yi (i≥ 3), mientras que la suma de ramas para la politomía (3,4,5,6) incluye 3 veces $(m-3)$ cada rama.

$$L_{xy} = \frac{1}{2(m-2)} \left[\sum_{i=3}^{m} (d_{1i} + d_{2i}) - (m-2) d_{12} - \frac{2}{m-3} \sum_{3 \le i < j} d_{ij} \right]$$



$$S_{12} = L_{1x} + L_{2x} + L_{xy} + \sum_{3 \le i < j} L_{ij}$$

$$L_{xy} = \frac{1}{2(m-2)}\left[\sum_{i=3}^{m}(d_{1i} + d_{2i}) - (m-2)d_{12} - \frac{2}{m-3}\sum_{3 \le i < j} d_{ij}\right]$$

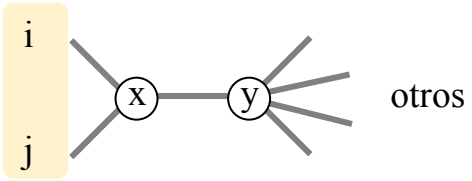
Reemplazando en S_{12} arriba obtenemos:

$$S_{12} = d_{12} + \frac{1}{2(m-2)}\sum_{i=3}^{m}(d_{1i} + d_{2i}) - \frac{1}{2}d_{12} - \frac{2}{2(m-3)(m-2)}\sum_{3 \le i < j}^{m} d_{ij} + \frac{1}{m-2}\sum_{3 \le i < j} d_{ij}$$

Simplificando:

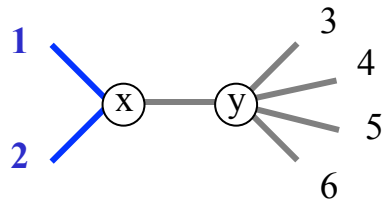
$$S_{12} = \frac{1}{2(m-2)}\sum_{i=3}^{m}(d_{1i} + d_{2i}) + \frac{1}{2}d_{12} + \frac{1}{m-2}\sum_{3 \le i < j} d_{ij}$$

Retomando el ejemplo



matriz de distancias

	1	2	3	4	5
2	9				
3	12	7			
4	15	10	5		
5	20	15	10	11	
6	16	11	6	7	8



$$S_{12} = \frac{1}{2(m-2)} \sum_{i=3}^{m} (d_{1i} + d_{2i}) + \frac{1}{2} d_{12} + \frac{1}{m-2} \sum_{3 \le i < j} d_{ij}$$

matriz S_{ij}

	1	2	3	4	5
2	**29.5**				
3	32.5	32.5			
4	33.0	33.0	32.0		
5	33.5	33.5	32.5	32.0	
6	33.5	33.5	32.5	32.0	30.5

$S_0 = 32.4$; $S_{ij} = 29.5$

Observamos: 1 y 2 forman el primer par, pero d_{12} no es la distancia menor en la matriz original.

¿A qué se debe que $S_{12} < S_0$?

Longitud de las ramas externas que llevan a los dos primeros vecinos

$$b_{ai} = \frac{1}{2(m-2)}[(m-2)d_{ij} + \sum_{k=1}^{m} d_{ik} - \sum_{k=1}^{m} d_{jk}]$$

Por simetría

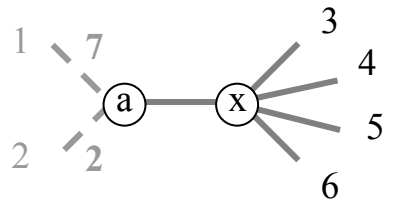
$$b_{aj} = \frac{1}{2(m-2)}[(m-2)d_{ij} + \sum_{k=1}^{m} d_{jk} - \sum_{k=1}^{m} d_{ik}]$$

Podemos reescribir estas distancias como:

$$b_{ai} = \frac{1}{2}d_{ij} + \frac{1}{2(m-2)}[\sum_{k=1}^{m}(d_{ik} - d_{jk})]$$

matriz de distancias

	1	**2**	**3**	**4**	**5**
2	9				
3	12	7			
4	15	10	5		
5	20	15	10	11	
6	16	11	6	7	8



$$b_{ai} = \frac{1}{2}d_{ij} + \frac{1}{2(m-2)}[\sum_{k=1}^{m}(d_{ik} - d_{jk})]$$

El primer término es la mitad de d_{ij}, al que sumamos un término basado en la diferencia de todas las distancias medidas desde el taxón i y desde el taxón j. Si esa diferencia es 0, $b_{ai} = b_{aj}$. De lo contrario, una de las ramas es mayor que la otra. El término que sumamos a una de las ramas es de igual magnitud y signo opuesto al que sumamos a la otra. Al igual que en fórmulas previas, las fracciones $\frac{1}{2}$ y $\frac{1}{2(m-2)}$ corrigen para el número de veces que pasamos por los segmentos o ramas de interés.

Más en general, todas las fórmulas para estimar la longitud de una o más ramas incluyen: a) una suma de distancias pareadas; b) a la cual le restamos las ramas que queremos excluir de dicha suma; y c) un factor de corrección con el cual dividimos el resultado de a y b por el número de veces que aparece la rama (o las ramas) de interés.

Longitud de la distancia entre el taxón a (sustituto del primer par de vecinos) y los restantes

Para cada uno de los taxones que quedan (en nuestro ejemplo 3, 4, 5 y 6, puesto que unimos a 1 y 2 el el primer ciclo), calculamos las distancias d_{ak}, $(k; 3 \leq k \leq m)$ como sigue:

$$d_{ak} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$$

Longitud de la distancia entre el taxón a (sustituto del primer par de vecinos) y los restantes

Para cada uno de los taxones que quedan (en nuestro ejemplo 3, 4, 5 y 6, puesto que unimos a 1 y 2 el el primer ciclo), calculamos las distancias d_{ak}, $(k; 3 \leq k \leq m)$ como sigue:

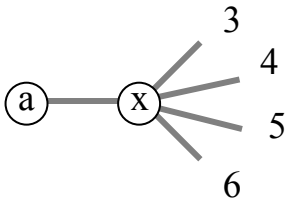
$$d_{ak} = \frac{1}{2}(d_{ik} + d_{jk} - d_{ij})$$

matriz de distancias inicial

	1	**2**	**3**	**4**	**5**
2	9				
3	12	7			
4	15	10	5		
5	20	15	10	11	
6	16	11	6	7	8

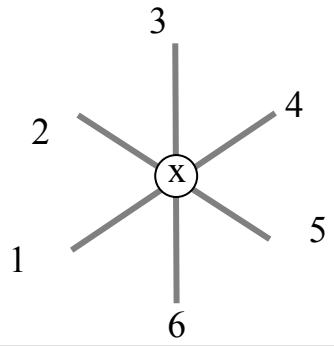
al final del primer ciclo

	a	**3**	**4**	**5**
3	5			
4	8	5		
5	13	10	11	
6	9	6	7	8

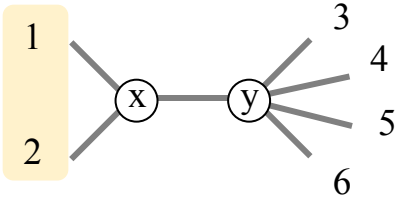


Método de unión de vecinos (NJ)

Paso 1: Comenzamos con una politomía (todos los taxa unidos a un único nodo interno).

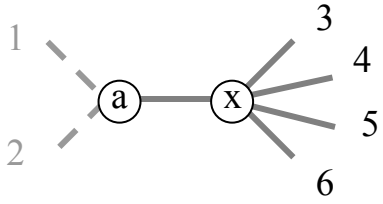


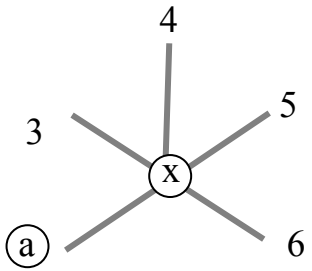
Paso 2: Elegimos un par de "vecinos" a unir: aquel que resulte en el árbol más corto.



(1,2), (3,4,5,6)
(1,3), (2,4,5,6)
….
(5,6) (1,2,3,4)

Paso 3: Reemplazamos dicho par por el nodo que los une (a). Para eso, volvemos a calcular las distancias entre todos los taxones (a y 3, 4, 5, 6 en el ejemplo).





- Vamos al Paso 1 con el árbol obtenido en el Paso 3.
- Repetimos el ciclo (1,2 3) hasta obtener un árbol completamente resuelto.
- Calculamos las longitudes de todas las ramas de dicho árbol.

matriz de distancias

	1	**2**	**3**	**4**	**5**
2	9				
3	12	7			
4	15	10	5		
5	20	15	10	11	
6	16	11	6	7	8

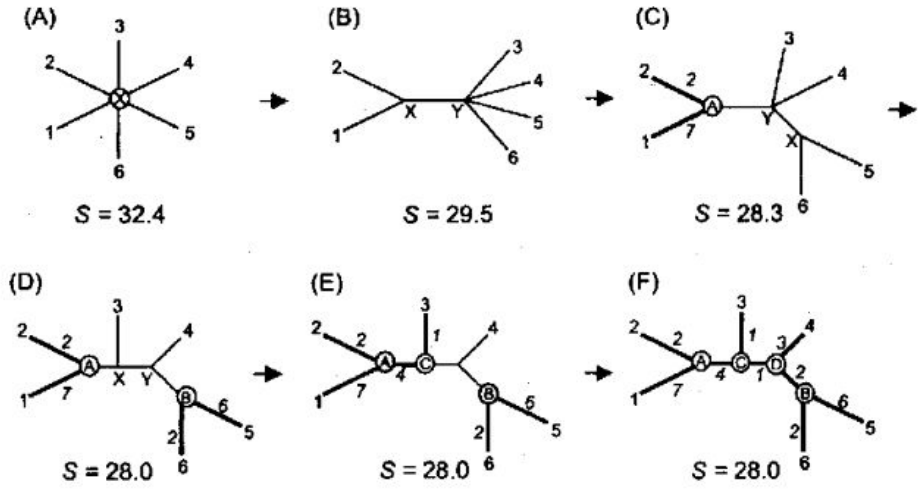


FIGURE 6.7. Illustration of the computational process in the neighbor-joining method.

- El ejemplo está basado en distancias sin error; en este caso (pero no en otros), las distancias finales (sumando ramas en el árbol) son exactamente las de la matriz inicial.
- Los OTUs que forman cada uno de los dos primeros pares ([1,2] y [5,6]) tienen distancias cortas, pero no necesariamente son las menores de la tabla.
- Las ramas terminales que llevan a 1 y 2 son notoriamente diferentes en longitud:
  - ¿qué nos dice esta observación?
  - ¿qué información en la tabla de distancias nos permite adelantar esta diferencia?

Algunos métodos de inferencia filogenética

Método	Variantes	Criterio de optimización	Uso de variación no observada
Parsimonia	Máxima parsimonia	Minimizar el número de pasos requeridos para obtener los datos	No
Distancias	- Evolución mínima	Minimizar los cambios requeridos para obtener las distancias estimadas entre taxa	Sí (incorporadas en las distancias)
	- Unión de vecinos (neighbor joining)	Una aproximación al árbol de evolución mínima	ídem
Inferencia estadística	Máxima Verosimilitud	Maximizar la verosimilitud de observar los datos, dado un árbol y un modelo de evolución molecular.	Sí (considerando todos los estados posibles en los nodos).
	Inferencia Bayesiana	…	

Algunas comparaciones

- Parsimonia vs. otros:
  - No se usan caracteres no observados.
  - Se trabaja con los datos originales, sin resumirlos como distancias pareadas.
  - No se utiliza un modelo de evolución (por ej., molecular)
  - El uso de los datos es idéntico, independientemente de la fuente (molecular, morfológica, ….). (Aunque existen algunas variaciones sobre este punto, por ejemplo aquellas que dan pesos distintos a distintos caracteres o cambios de estado).
- Inferencia estadística (verosimilitud, bayesiana) vs. distancias:
  - En común: modelos de evolución molecular y distancias asociadas.
  - Diferencias: criterio y forma de optimización.