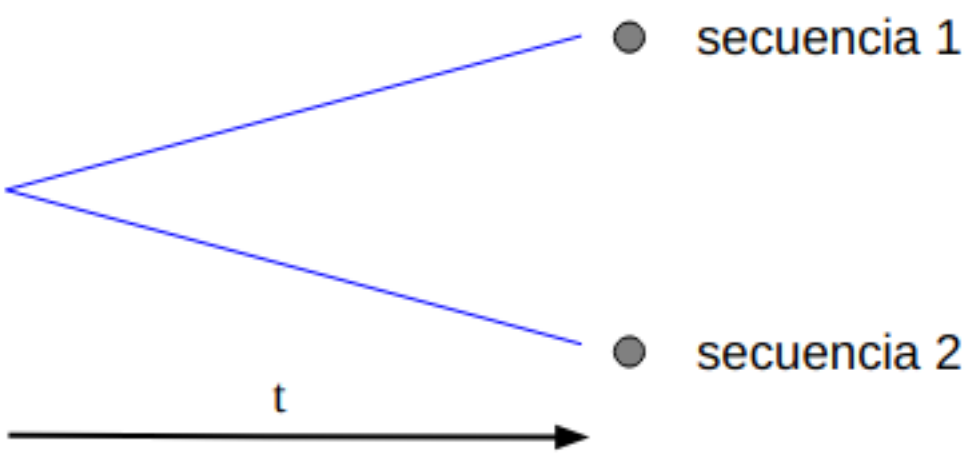
Neighbor joining v 2

Enrique Lessa

2024-08-28

Distancia de Jukes-Cantor

Vamos a considerar la divergencia de dos secuencias a partir de su ancestro para deducir cómo cambian a lo largo del tiempo.



Para eso, tenemos que definir el proceso evolutivo, en este caso con un modelo sencillo: el de Jukes-Cantor. El mismo está caracterizado por la siguiente matriz de transición, que consideramos válida para todos los sitios en una secuencia de ADN. Empezamos por un modelo en tiempos discretos. La matriz dice que, si un sitio está en un estado cualquiera (una de las letras en el margen izquierdo de la matriz), la probabilidad de pasar, en una unidad de tiempo, a cualquier otro estado es la misma (α). La probabilidad de permanecer en el mismo estado es, por tanto, $1 - 3\alpha$.

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

matriz

Un esquema de cómo se estima la distancia de Jukes-Cantor (d_{JK}) a partir de la distancia observada (proporción de sitios diferentes p) entre dos secuencias:

- Llamamos q_t a la fracción de sitios idénticos en las dos secuencias en el tiempo t. $q_t = 1 - p_t$.
- La probabilidad de que un sitio, en cualquiera de las dos secuencias, sufra una sustitución (pase a un estado diferente) es $k = 3\alpha$. k es la tasa de sustitución por sitio y por unidad de tiempo.
- Los sitios que son idénticos en t se mantendrán iguales en una fracción $\left(1-k\right)^2$; despreciando el término k^2, dicha fracción es $1 - 2k$. [¿Por qué esta aproximación es razonable? Pensar en qué miden los términos $2k$ y k^2.]
- Los sitios que son diferentes en t serán iguales en $t + 1$ en una fracción $2k\left(1-k\right)/3$, que aproximamos a $2k/3$.

Como resultado:

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

El cambio neto es:

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

Pasando a tiempo continuo, tenemos:

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

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

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

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:

$$E\left(d_{JK}\right) = 2kt$$

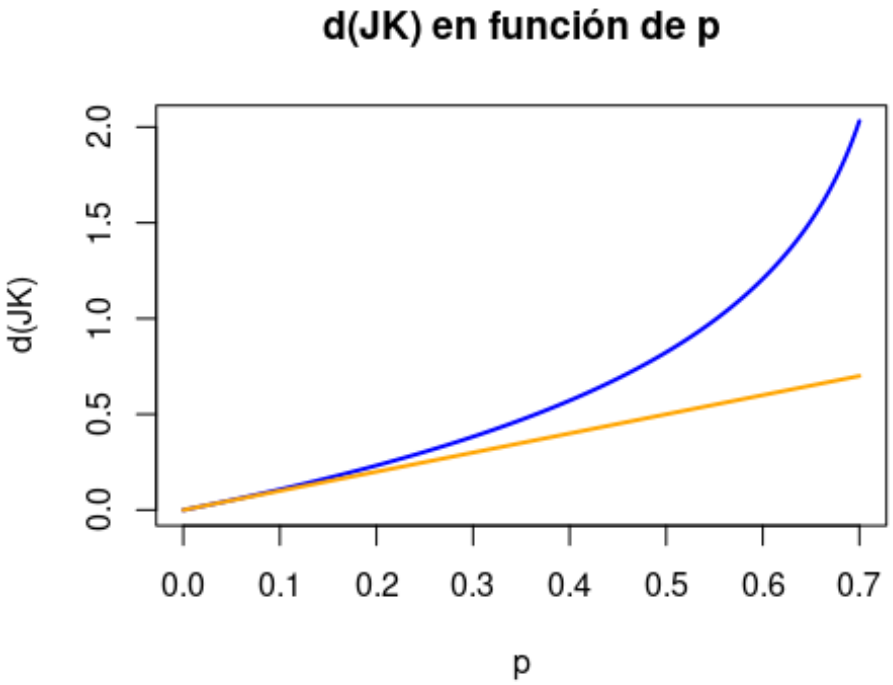
Esta es una k de cada una de las ramas que van desde el ancestro común hasta las secuencias que estamos comparando.

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

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

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.

```
curve(  -(3/4)*log(1-(4/3)*x), from=0, to=.7, n=300, xlab="p",
ylab="d(JK)",
             col="blue", lwd=2, main="d(JK) en función de p"  )
curve((x), add=TRUE, n=300, lwd=2, col="orange")
```



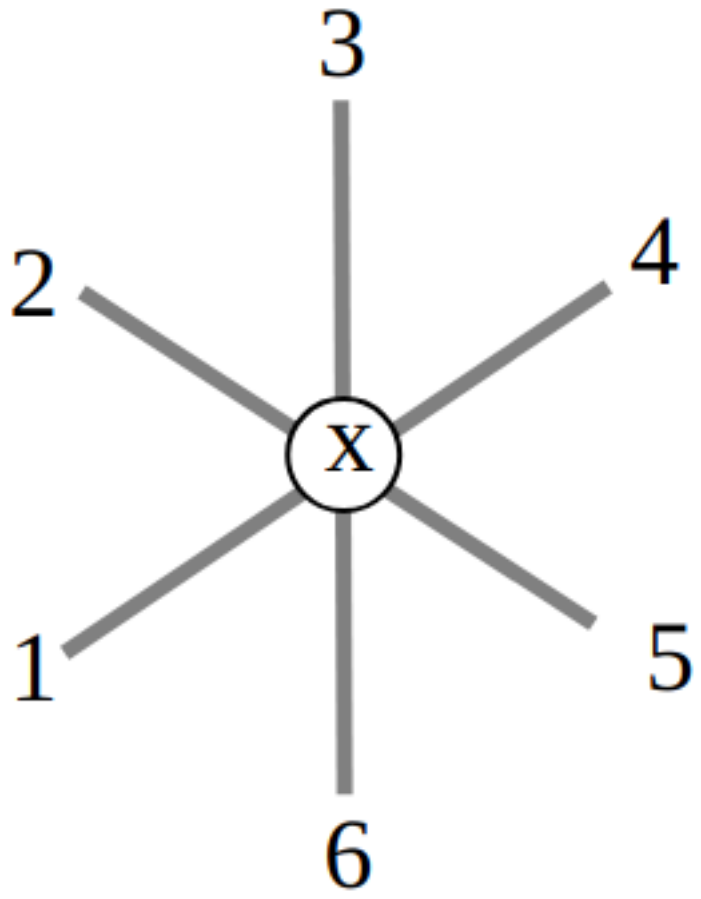
Podemos extraer varias enseñanzas de estas gráficas:

1) Por debajo de cierto nivel de divergencia observada (del orden de 10%), las diferencias entre p y d_{JK} son despreciables.

2) Notemos que esperamos que 2 secuencias completamente al azar diverjan en 75% de sus sitios. Por tanto, valores empíricos de p cecanos a 0.75 (por encima de 50-60%) son compatibles con distancias muy diferentes. En esa zona, nuestras estimaciones de d_{JK} dejan de ser confiables.

Unión de vecinos

Trabajaremos un ejemplo con 6 taxones ($m = 6$).



Árbol inicial para 6 taxones

El punto de partida es una tabla de distancias entre pares de taxones como la siguiente:

matriz de distancias

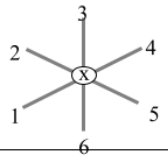
	1	2	3	4	5
2	9				
3	12	7			
4	15	10	5		
5	20	15	10	11	
6	16	11	6	7	8

matrix de distancias entre pares de taxones

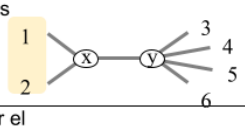
Esquemáticamente, el procedimiento de unión de vecinos es el siguiente:

Método de unión de vecinos (NJ)

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

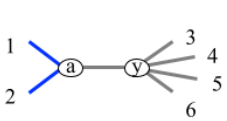


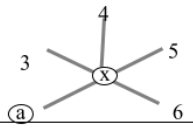
Paso 2: Elegimos un par de "vecinos" a unir: aquel que resulte en el árbol más corto.



(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.

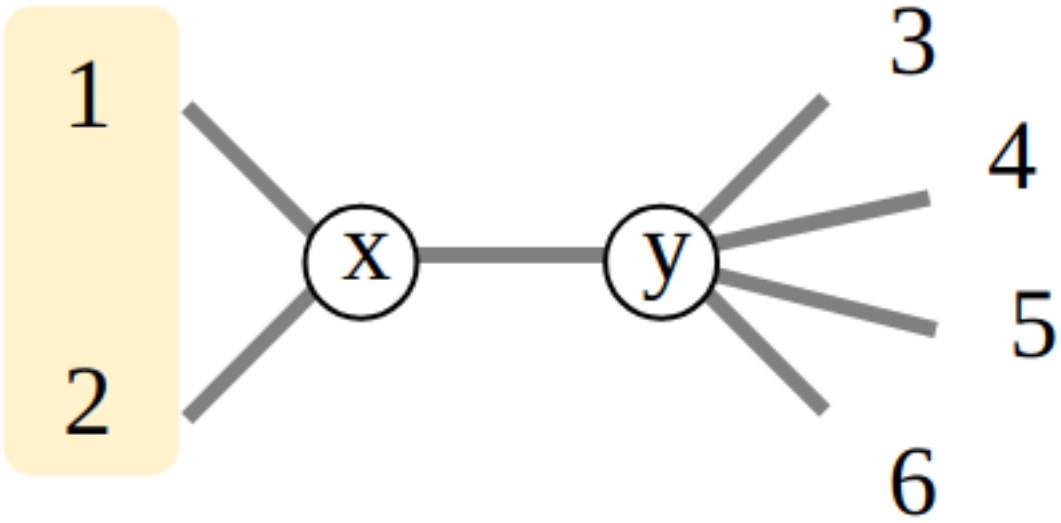
16

Algoritmo de NJ: unión de vecinos

Elección del primer par de vecinos

Enumeramos todos los pares candidatos a ser los primeros dos "vecinos" $(1,2)$, $(1,3)$, ... $(m-1,m)$ 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.



Árbol uniendo el par12

La longitud total de este árbol es:

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

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

Y tenemos una fórmula para el tercer término: 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 $y\,i$ (i $\geq$ 3), mientras que la suma de ramas para la politomía (3,4,5,6) incluye 3 veces $(m-3)$ cada rama.

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

Reemplazando en S_{12} arriba obtenemos:

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

Simplificando:

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

Vamos a repetir el procedimiento para obtener los valores de S_{ij} para todos los pares de taxones, y podemos resumir los resultados en una tabla como esta:

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

Matriz de longitudes de árboles S_{ij}

Notamos que el árbol más corto entre los que resuelven el primer par es aquel que une a los taxones 1 y 2. Notamos también, comparando con la matriz original de datos más arriba, que ese par no es aquel que presenta la menor distancia. Debería ser ese par en el caso en que las distancias son proporcionales al tiempo de separación. Obviamente, aunque no sabemos los detalles, esto no se cumple en este caso. El par (3,4) es el que tiene menor distancia, y no es el seleccionado en el primer paso.

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

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

Por simetría

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

Podemos reescribir estas distancias como:

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

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 rama que une 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:

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

Distancias al final del primer ciclo

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

matrixtaxa

A continuación se reproduce un resumen esquemático de la resolución de este ejemplo tomado del libro de Nei & Kumar, 2000. Molecular Evolution and Phylogenetics, Oxford Univ. Press.

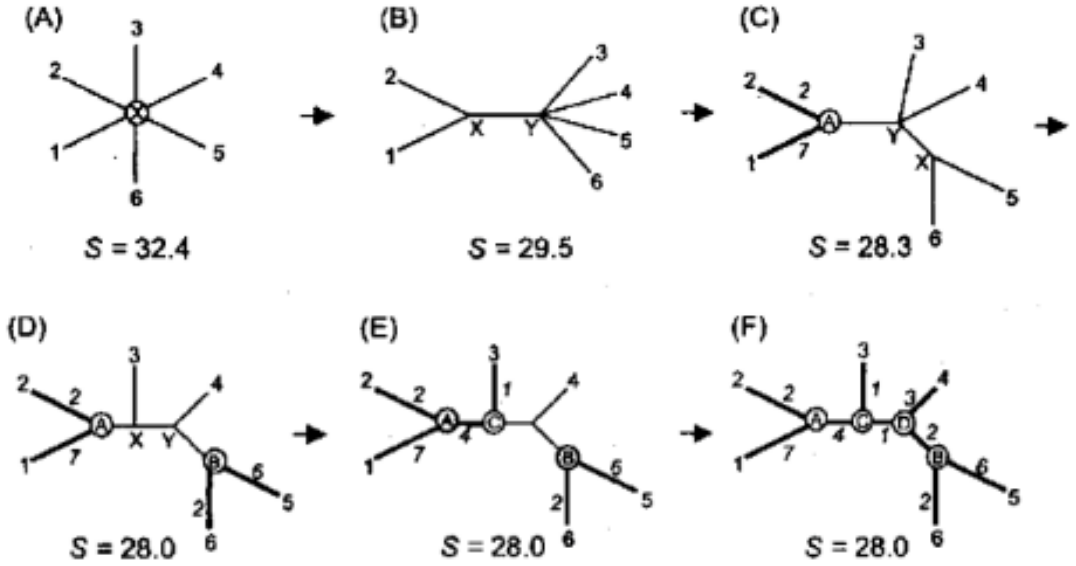


FIGURE 6.7. Illustration of the computational process in the neighbor-joining method.

En (A) vemos el árbol de partida y su longitud. En (B) el árbol que define el primer par (1,2) y su longitud. En (C) están calculadas las longitudes de las ramas A1 y A2, y también el nuevo par elegido (5,6) siguiendo el método.