I CURSILLO DE FILOGENIA PARA USUARIOS

SESIÓN 3 – Máxima Verosimilitud y Análisis Bayesiano

MÁXIMA VEROSIMILITUD

Trata de encontrar el árbol más probable.

La máxima verosimilitud calcula la probabilidad de obtener los datos (D, la matriz) dada una determinada hipótesis (H, el árbol y el modelo de sustitución): cuál es la probabilidad de que, habiendo obtenido un determinado árbol, éste sea el resultado de nuestros datos.

El mejor árbol es el que tenga mejor P.

$$P(D|H)$$

MÁXIMA VEROSIMILITUD

Una consideración “filosófica”:

La ecuación de verosimilitud **no** es la probabilidad de que la hipótesis sea correcta en términos absolutos, sino para nuestros datos.

$$P(D|H)$$

MÁXIMA VEROSIMILITUD

Sequence 1 **C C A T**

Sequence 2 **C C G T**

$$\pi = [0.1, 0.4, 0.2, 0.3]$$

Cálculo:

$$\boldsymbol{P} = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

$$L_{(Seq.1 \to Seq.2)} = \pi_C P_{C \to C} \pi_C P_{C \to C} \pi_A P_{A \to G} \pi_T P_{T \to T}$$

$$0.4\text{x}0.983\text{x}0.4\text{x}0.983\text{x}0.1\text{x}0.007\text{x}0.3\text{x}0.979$$

$$= 0.0000300$$

$$lnL_{tree:Seq_1 \to Seq_2} = -10.414$$

MÁXIMA VEROSIMILITUD

¿Por qué se expresa en logaritmo neperiano?

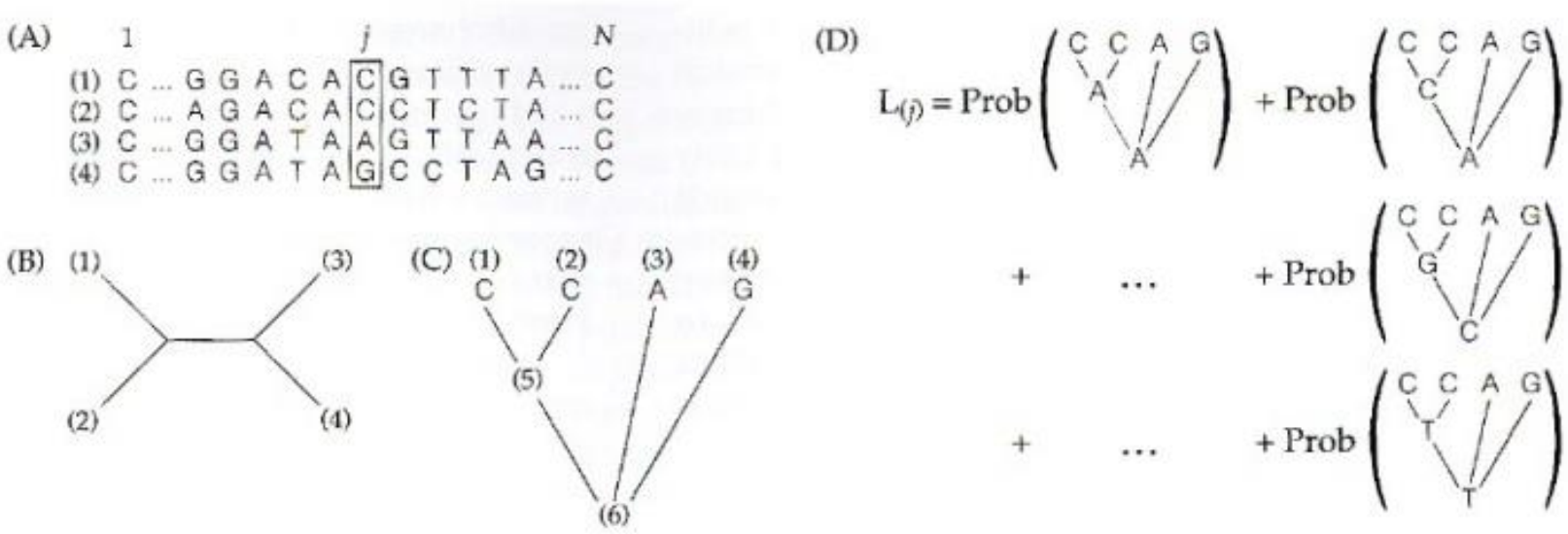
Son valores muy pequeños: 0,0…(miles de ceros)…01456

Los ordenadores no pueden almacenar números tan pequeños en su memoria, al igual que tampoco pueden trabajar con números muy grandes. Por ejemplo, el número más pequeño con el que puede trabajar mi ordenador es e^{-1021}

Además hay problemas de precisión (como lo que pasa en las calculadoras).

MÁXIMA VEROSIMILITUD

Cálculo:



Se repite para cada posición y se van multiplicando (o sumando, si trabajamos con logaritmos).

Se puede añadir además la longitud de las ramas (multiplicando).

MÁXIMA VEROSIMILITUD

Al igual que para la MP, tenemos el problema del número de árboles: no podemos evaluarlos todos.

Algoritmo de "**hill climbing**":

- Se toma un árbol de partida, calculado mediante NJ
- Se calcula su verosimilitud
- A partir de ahí se van introduciendo cambios (topología, longitud de las ramas, parámetros del modelo…). Cada vez que se cambia algo, se recalcula la verosimilitud: si es menor que la del árbol previo, se descarta; si es mayor, descartamos el árbol previo y nos quedamos con el nuevo.

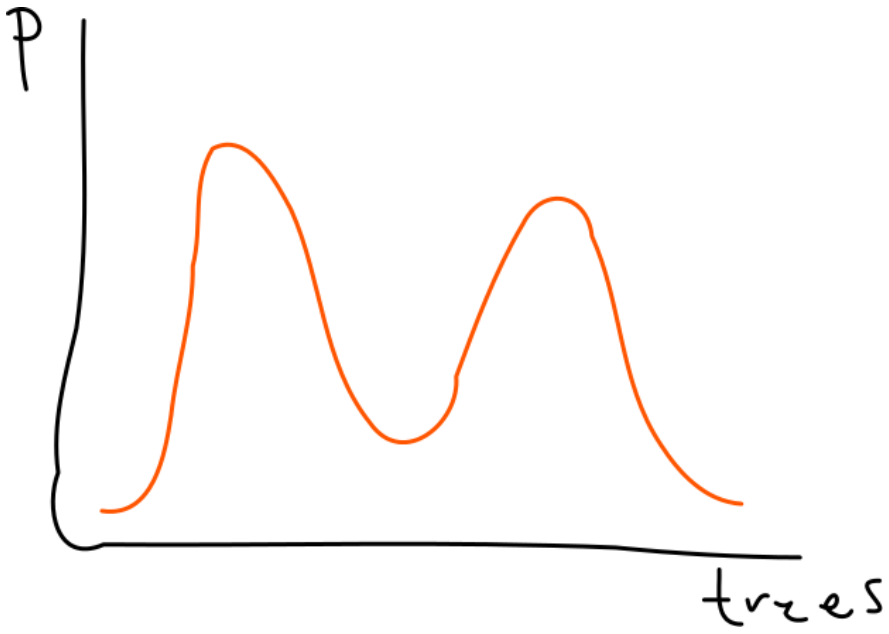
MÁXIMA VEROSIMILITUD

Ventajas:

- Rápido
- Fiable

Problemas:

- Fenómeno de atracción de ramas cortas
- No considera soluciones alternativas



Es mejor hacer ML que NJ, ya que duran lo mismo y el NJ no tiene en cuenta posibles fenómenos (como homologías) que afectan al resultado.

MÁXIMA VEROSIMILITUD

Dado que cambiando los parámetros de partida obtenemos valores de lnL diferentes, podemos usar este método para calcular cuáles son los valores óptimos de dichos parámetros.

Algunos ejemplos:

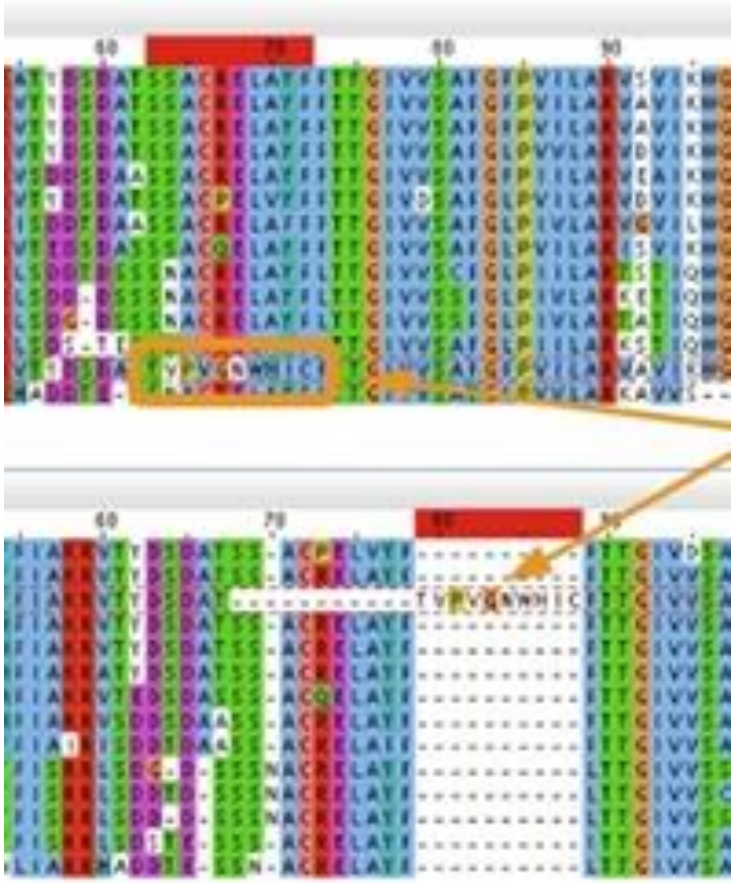
1. **¡Modeltest:** Hace análisis paralelos usando un modelo de sustitución nucleotídica diferente en cada uno. El que tenga mayor lnL será el más adecuado para nuestros datos.
2. **Test de topologías:** Si nos da un árbol inesperado como el más probable, podemos repetir el análisis fijando la topología que esperábamos y comparar los valores de lnL de ambos. Si la diferencia es significativa concluimos que nuestra hipótesis inicial realmente es incorrecta; si no lo es, podemos decir que no puede ser descartada.

ALINEAMIENTO

Métodos **clásicos** (ClustalW)

Métodos **iterativos** (MUSCLE)

Métodos **guiados por filogenias** (PAGAN)



MÁS PARÁMETROS

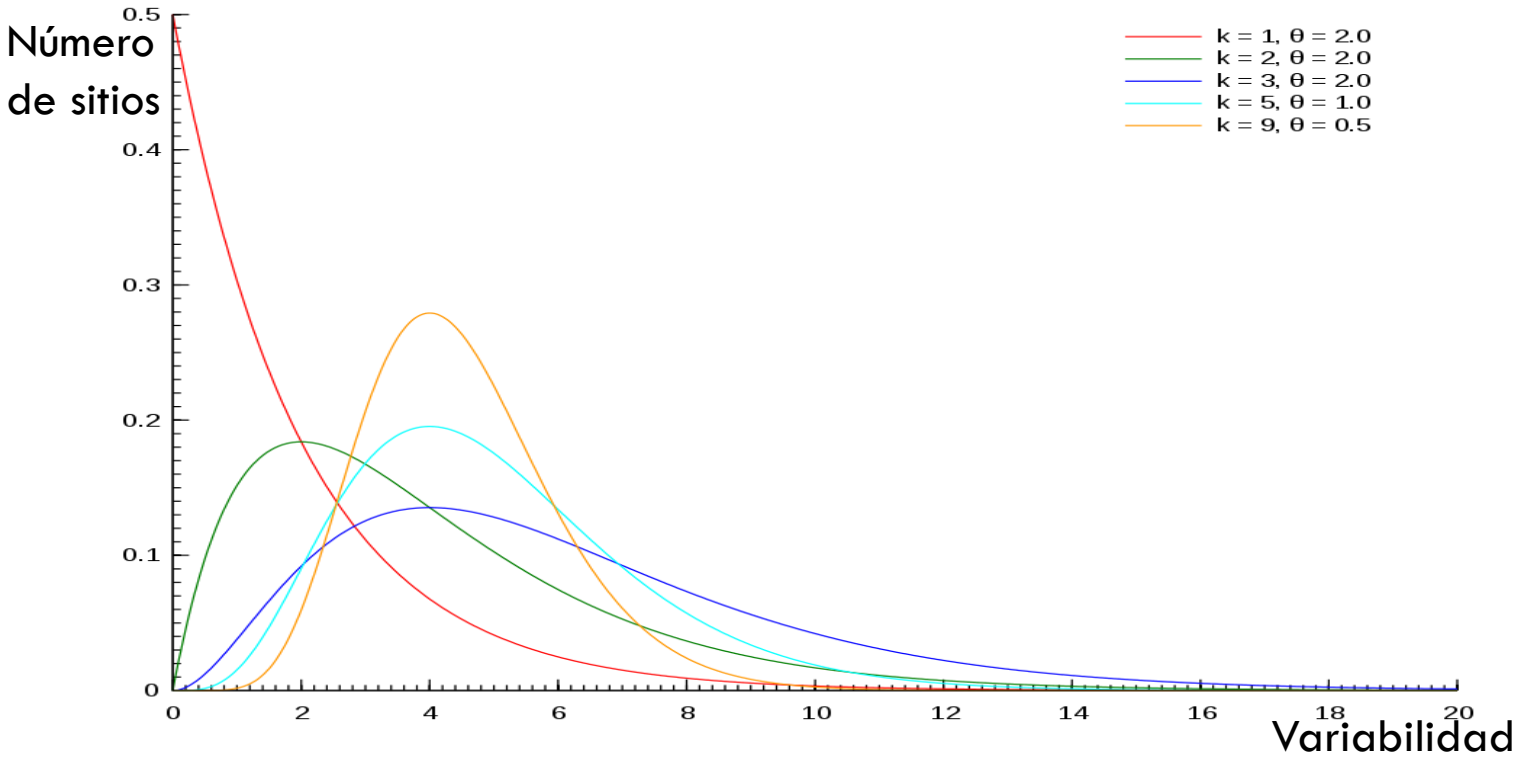
(modelo) + I + G

I (invariants): Proporción de sitios invariantes

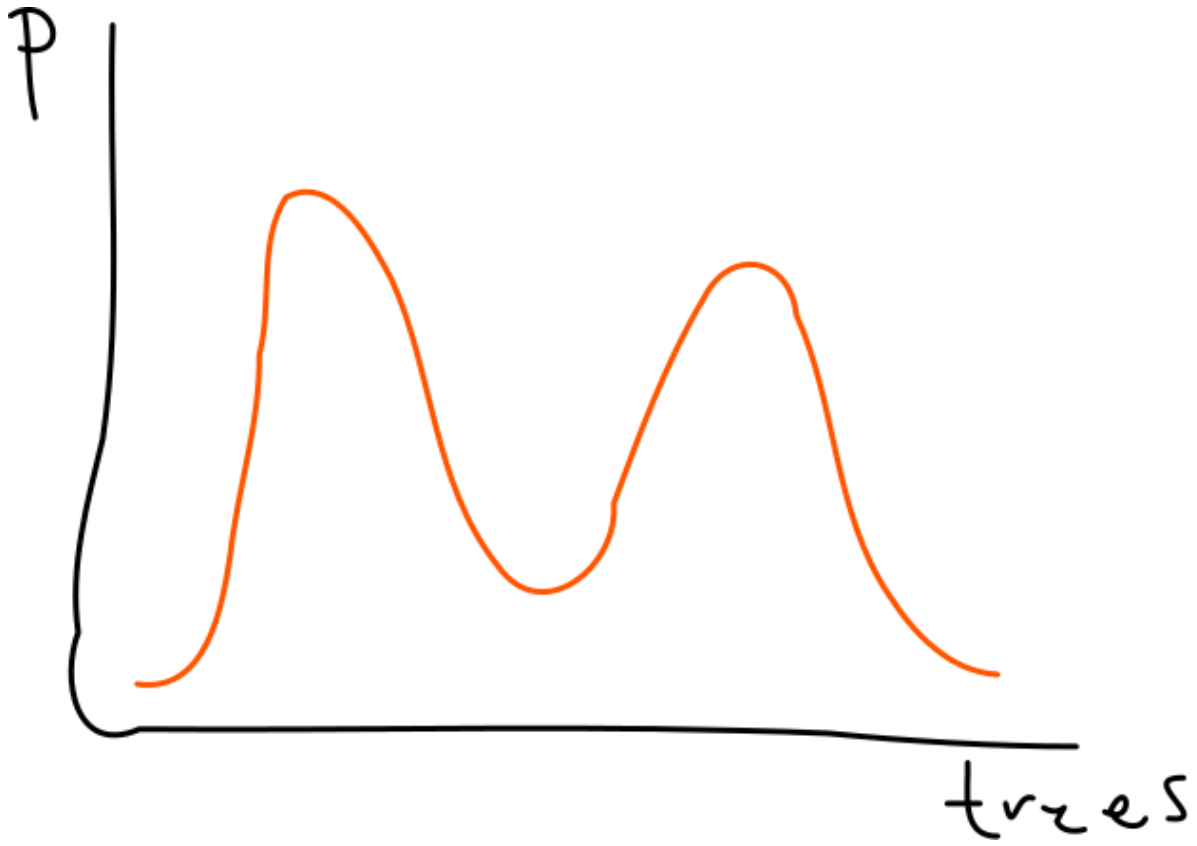
G (gamma): Distribución de la variabilidad a lo largo de los sitios

MÁS PARÁMETROS

Distribución gamma:



INFERENCIA BAYESIANA



INFERENCIA BAYESIANA

La Máxima Verosimilitud trataba de calcular la probabilidad de obtener los datos (D, la matriz) dada una determinada hipótesis (H, el árbol y el modelo de sustitución).

$$P(D|H)$$

La Inferencia Bayesiana calcula la probabilidad de cierta hipótesis (H, el árbol) dados unos ciertos datos de partida (D, la matriz).

$$P(H|D)$$

INFERENCIA BAYESIANA

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

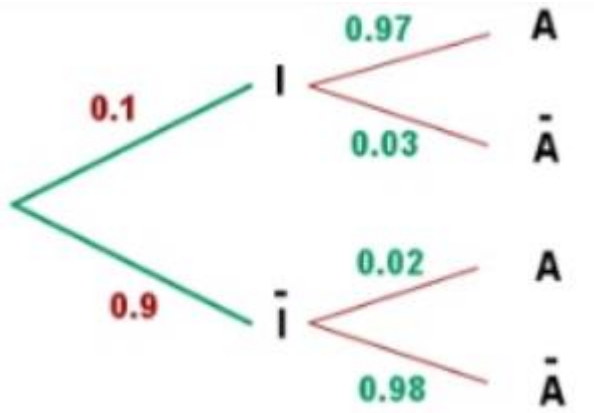
INFERENCIA BAYESIANA

Es una probabilidad "a posteriori".

Ejemplo: Hay una fábrica con alarma. La probabilidad de que haya un accidente es 0,1. La probabilidad de que la alarma suene en caso de accidente es 0,97. La probabilidad de que se dispare accidentalmente es 0,02. Acaba de sonar la alarma, ¿cuál es la probabilidad de que no haya habido ningún accidente?

$$P(nI|A) = \frac{P(A|nI) \cdot P(nI)}{P(A)} =$$

$$= \frac{0{,}02 \cdot 0{,}9}{(0{,}1 \cdot 0{,}97) + (0{,}9 \cdot 0{,}02)} = 0{,}157$$



INFERENCIA BAYESIANA

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Likelihood

Probabilidad "a priori" del árbol

No se sabe cómo calcularla, así que se asume que es igual a:

$1/N_{árboles\ posibles}$

Probabilidad de los datos

Imposible de calcular. Se elimina al usar las MCMC.

INFERENCIA BAYESIANA

MCMC (Markov chain Monte Carlo)

1. Se crea un espacio multidimensional de todos los árboles posibles.
2. Empezamos en un árbol al azar.
3. Andamos un "paso".
4. Tomamos el nuevo árbol y calculamos:

$$R = \frac{Bayes(T_{nuevo})}{Bayes(T_{anterior})} = \frac{\frac{P(H_n) \cdot P(D|H_n)}{P(D)}}{\frac{P(H_a) \cdot P(D|H_a)}{P(D)}} = \frac{P(D|H_n)}{P(D|H_a)}$$

5. Si R>1, seguimos desde donde estamos. Si R<1, volvemos al árbol anterior.

INFERENCIA BAYESIANA

Además de esta “cadena fría”, tenemos dos “cadenas calientes” que van explorando árboles más “lejanos” en búsqueda de otros picos de probabilidad.

Si una “cadena caliente” encuentra un árbol con probabilidad más alta, se “enfría” y recoge el testigo. La antigua “cadena fría” se “calienta” y empieza a explorar árboles lejanos.

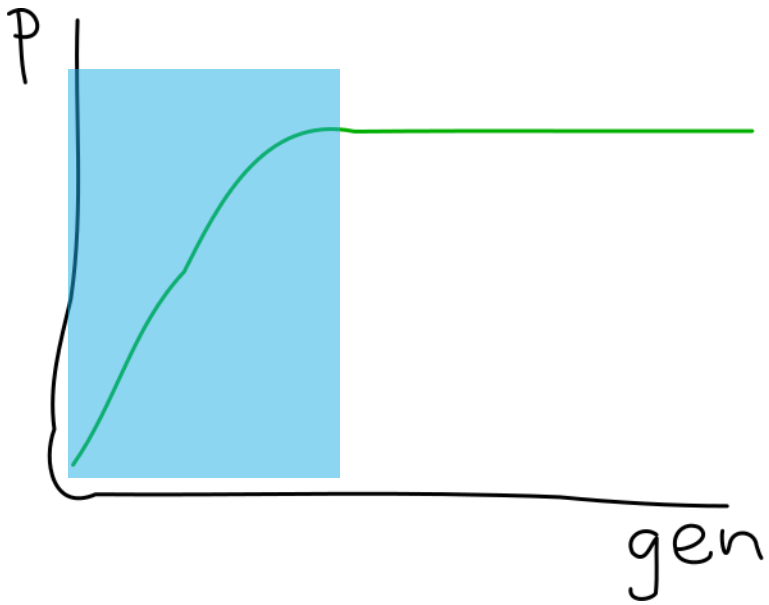
INFERENCIA BAYESIANA

```
85500 -- [-3279.417] (-3282.465) (-3280.517) (-3278.937)
86000 -- (-3281.087) (-3284.846) [-3275.492] (-3274.322)
86500 -- [-3283.089] (-3286.798) (-3279.602) (-3282.224)
87000 -- (-3278.794) [-3280.669] (-3283.353) (-3277.553)
87500 -- (-3276.343) (-3278.018) (-3278.278) [-3276.394]
88000 -- [-3282.651] (-3280.952) (-3286.720) (-3276.827)
88500 -- (-3274.662) (-3289.903) [-3279.075] (-3278.377)
89000 -- (-3281.422) (-3277.367) [-3275.465] (-3282.392)
89500 -- [-3273.179] (-3280.484) (-3274.539) (-3285.715)
90000 -- [-3274.872] (-3278.899) (-3277.294) (-3292.835)
```

INFERENCIA BAYESIANA

Cada tanto pasos, la cadena fría "muestrea" el árbol en el que se encuentra en ese momento, guardándolo en la memoria.

Al final se hace un árbol consenso con todos los árboles obtenidos, eliminando los primeros (**burn-in**).



INFERENCIA BAYESIANA

Es lento y requiere una considerable potencia de cálculo.

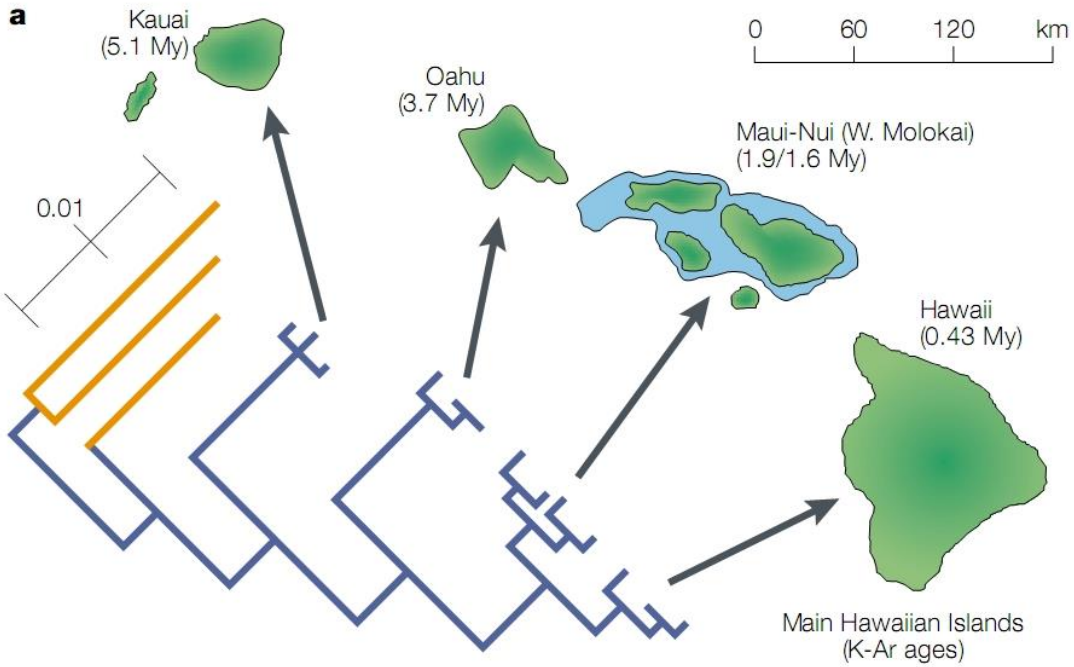
Es poco factible realizar *bootstrap* en IB con los ordenadores actuales. Un análisis típico dura horas, o incluso días; un *bootstrap* con 10000 réplicas duraría años (9 años para un análisis de 8 horas).

El soporte viene dado por la **probabilidad posterior** de cada nodo, que se calcula durante el análisis.

Ojo: NO es comparable con el *bootstrap* (BPP aceptable si >90%).

Pero es más rápido (a veces) y fiable que ML con *bootstrap*.

DATACIÓN



Puntos de calibración:

1. Fósiles (lower bound)
2. Biogeografía
   1. Vicarianza (fixed)
   2. Archipiélagos (upper bound)

Reloj molecular (tasa de sustitución)

1. Estricto
2. Relajado

¿EL FUTURO?