Modelos evolutivos y máxima verosimilitud

Datos Moleculares II

Nelson R. Salinas

Septiembre 19, 2020

Máxima Verosimilitud
Maximum Likelihood

- Método de estimación de parámetros basado en probabilidades.
- Popularizado por Ronald Fisher.
- Requiere un modelo estadístico que conceptualiza el proceso.

Ronald Fisher (1890-1962)

$$L(h) = p(D|\theta)$$

donde h es la hipótesis que se evalua, D los datos observados y θ los parámetros del modelo. En cierta manera es similar en términos generales con la minimización de costo de una función objetiva (o parsimonia):

$$C(h) = f(D, \theta)$$

Donde *C(h)* es el costo de la hipótesis h y *f()* es la función objetiva, que también tiene sus propios parámetros.

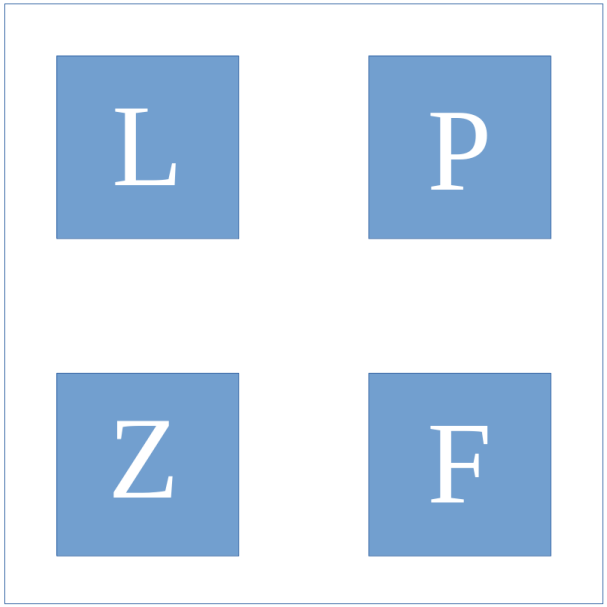
La diferencia entre máxima verosimilitud y parsimonia en la valoración de hipótesis radica en la aplicación del principio de probabilidad:

- $0 \leq L(h) \leq 1$
- La suma de $L(h)$ para todas las hipótesis posibles h es 1.

Ambos supuestos no aplican en parsimonia.

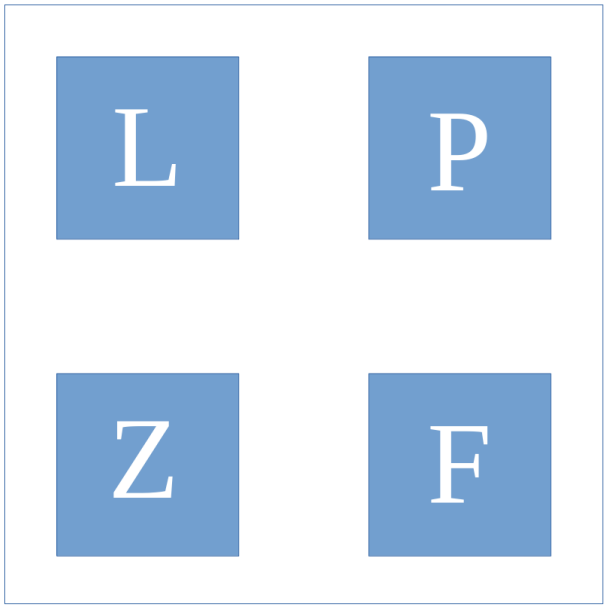
Ejemplo de parsimonia

- Una panadería, una zapatería, una floristería y una librería.
- Un individuo visita algunos negocios y se quiere estimar la secuencia de visitas observando los elementos comprados.



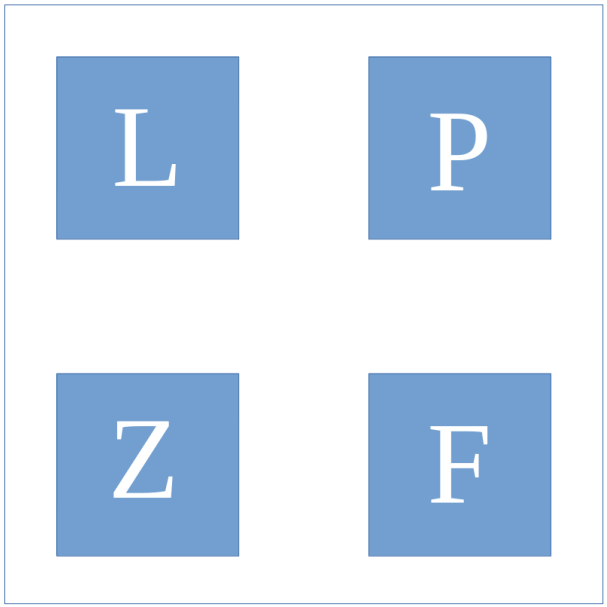
Ejemplo de parsimonia

- La función de coste de parsimonia: $C(h) = -O + N + F$
  - O = observado e incluido en h.
  - N = no observado e incluidos en h.
  - F = observado y no incluido en h.



Ejemplo de parsimonia

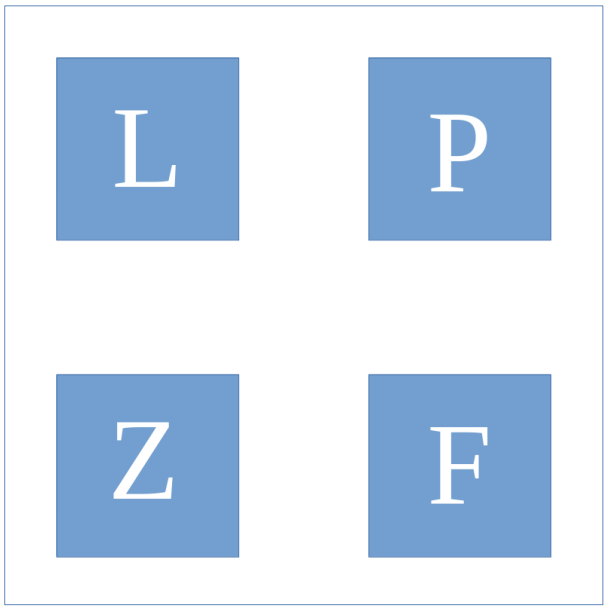
- La función de coste de parsimonia: $C(h) = -O + N + F$
- La mejor hipótesis maximiza C.
- Si se observan un pan, un racimo de flores y un libro, la hipótesis panadería - floristería - librería es mejor evaluada que panadería - floristería - zapatería (-3 vs. 0).



Ejemplo de ML

- Modelo estadístico gobierna las probabilidades de transición entre locales.

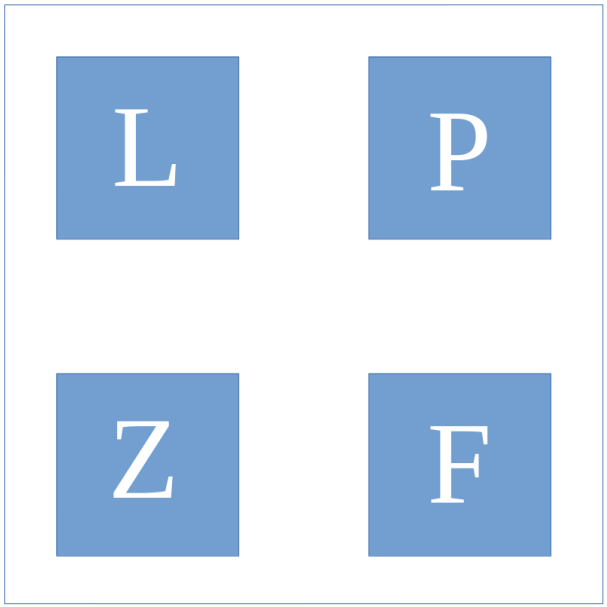
de\a	F	L	P	Z
F	0.25	0.25	0.25	0.25
L	0.25	0.25	0.25	0.25
P	0.25	0.25	0.25	0.25
Z	0.25	0.25	0.25	0.25



Ejemplo de ML

- $L(h) = p(D|\theta)$
- Datos: pan - flores - libro.
- Hipótesis: P-F-L.
- $(0.25 \times 1) \times (0.25 \times 1) \times (0.25 \times 1) = 0.015625$

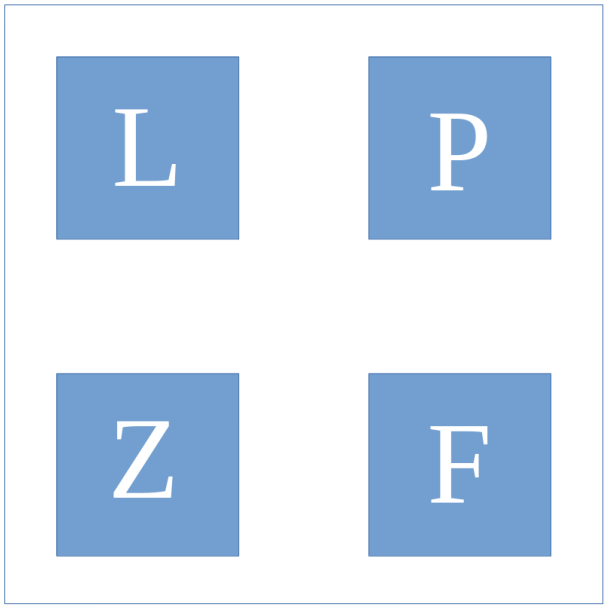
de\a	F	L	P	Z
F	0.25	0.25	0.25	0.25
L	0.25	0.25	0.25	0.25
P	0.25	0.25	0.25	0.25
Z	0.25	0.25	0.25	0.25



Ejemplo de ML

- $L(h) = p(D|\theta)$
- Datos: pan - flores - libro.
- Hipótesis: P-F-Z.
- $(0.25 \times 1) \times (0.25 \times 1) \times (0.25 \times 0) = 0$

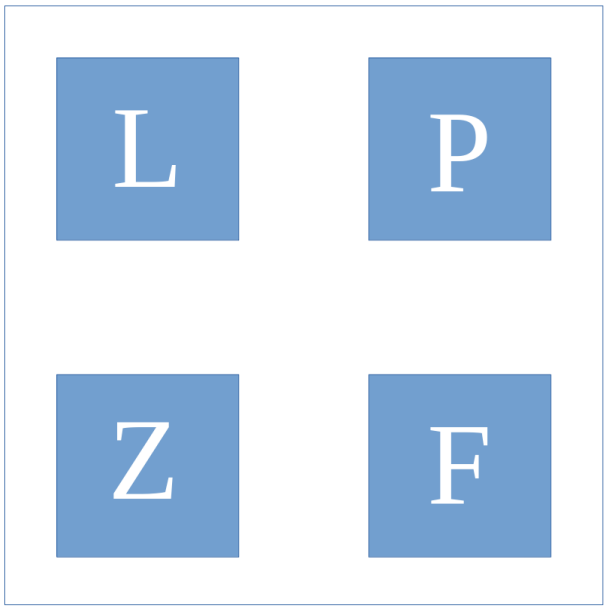
de\a	F	L	P	Z
F	0.25	0.25	0.25	0.25
L	0.25	0.25	0.25	0.25
P	0.25	0.25	0.25	0.25
Z	0.25	0.25	0.25	0.25



Ejemplo de ML

- $L(h) = p(D|\theta)$
- Datos: pan - flores - zapato.
- Hipótesis: P-F-Z.
- $(0.25 \times 1) \times (0.25 \times 1) \times (0.5 \times 1) = 0.03125$

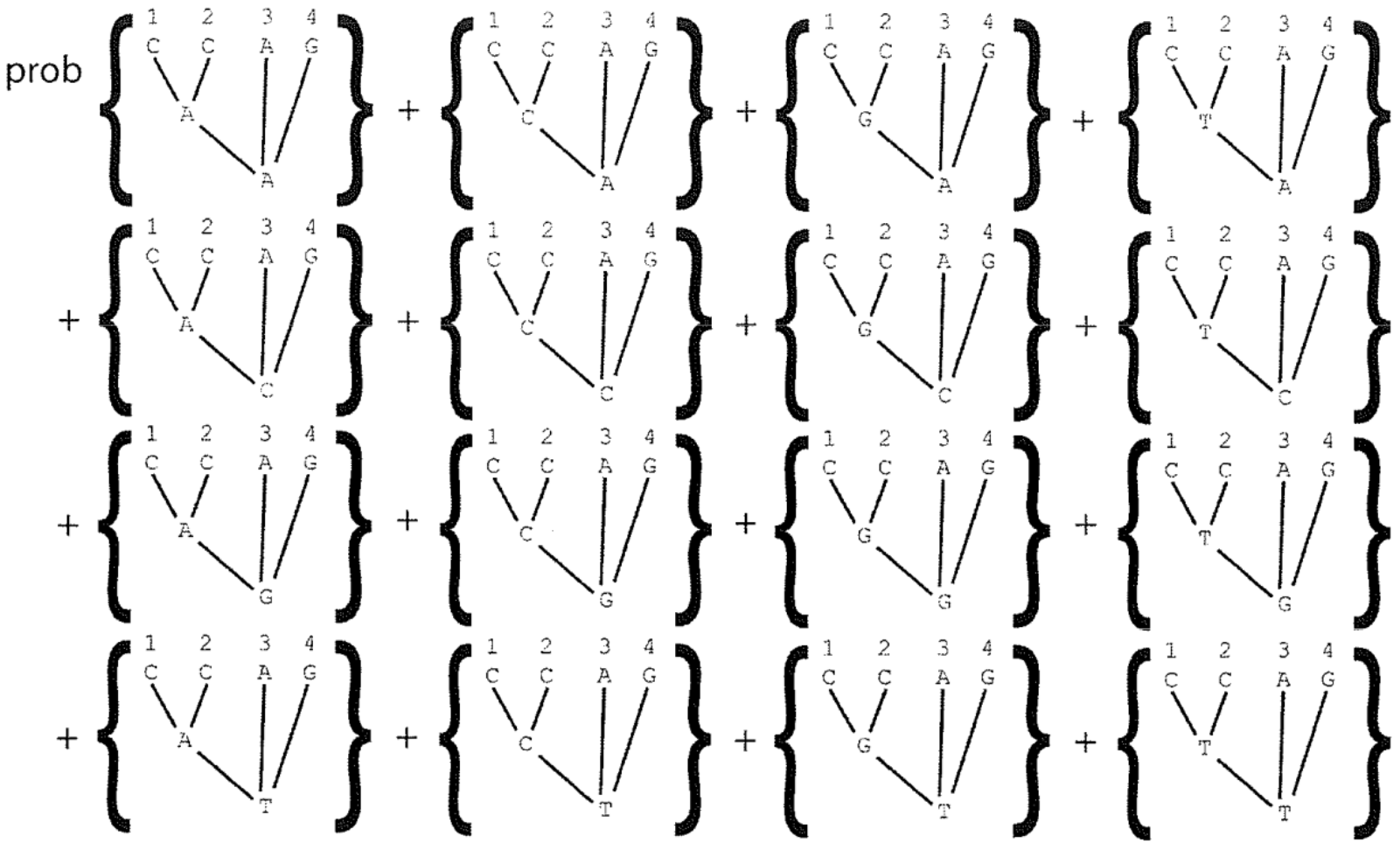
de\a	F	L	P	Z
F	0.0	0.25	0.25	0.5
L	0.25	0.0	0.25	0.5
P	0.25	0.25	0.0	0.5
Z	0.2	0.2	0.2	0.4



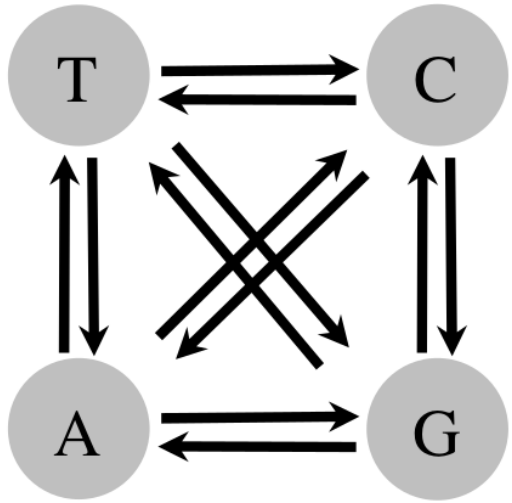
- Inferencia filogenética también utiliza matrices de substitución.
- Probabilidad de un árbol = suma de las posibles combinaciones de nucleóticos en los nodos.

	A	C	G	T
A	0.0	0.25	0.25	0.5
C	0.25	0.0	0.25	0.5
G	0.25	0.25	0.0	0.5
T	0.2	0.2	0.2	0.4

Máxima Verosimilitud
Maximum Likelihood

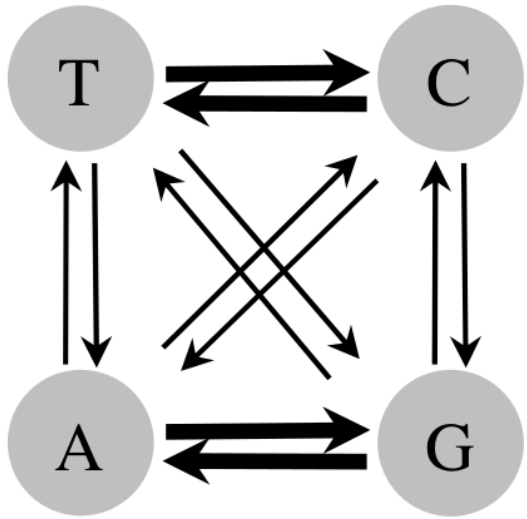


- Existen multiples tipos de modelos de substitución.
- Diferencias en el número de parámetros.
- Parametros: tasas de substitución entre nucleótidos, frequencia en estado estacionario y reversibilidad.



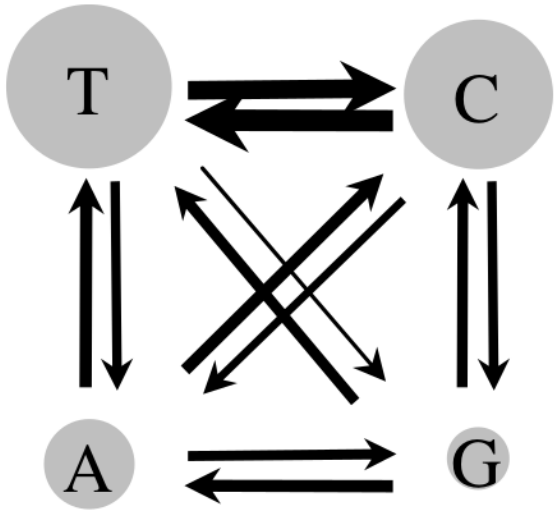
JC69 (Jukes-Cantor 1969)

- Existen multiples tipos de modelos de substitución.
- Diferencias en el número de parámetros.
- Parametros: tasas de substitución entre nucleótidos, frequencia en estado estacionario y reversibilidad.



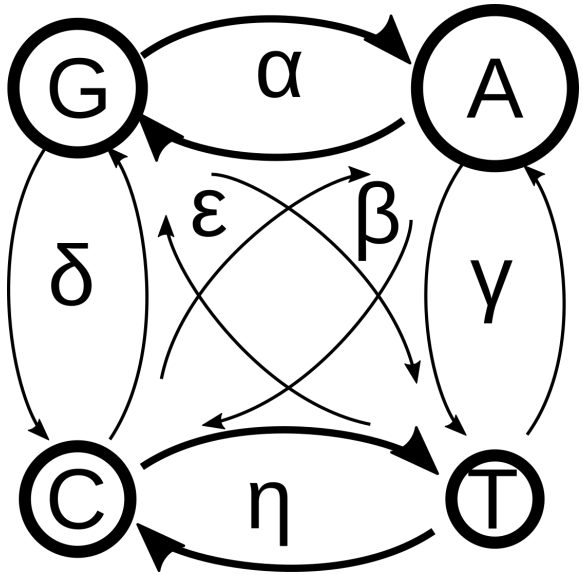
K80 (Kimura 1980)

- Existen multiples tipos de modelos de substitución.
- Diferencias en el número de parámetros.
- Parametros: tasas de substitución entre nucleótidos, frequencia en estado estacionario y reversibilidad.



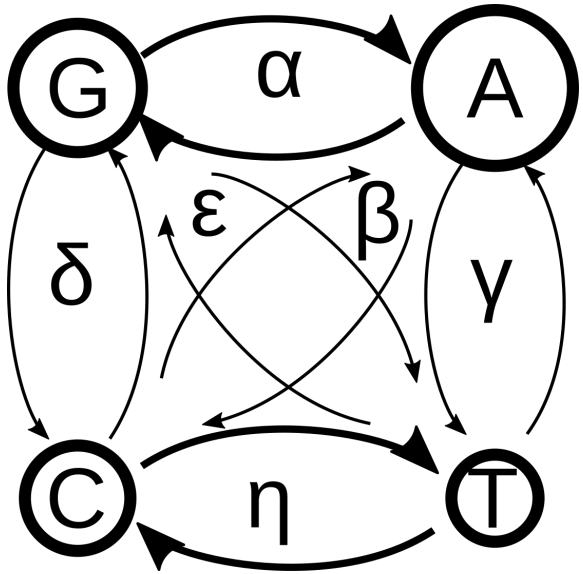
HKY85 (Hasegawa et al. 1985)

- Existen multiples tipos de modelos de substitución.
- Diferencias en el número de parámetros.
- Parametros: tasas de substitución entre nucleótidos, frequencia en estado estacionario y reversibilidad.



GTR (Tavaré 1986)

- GTR es el modelo más ampliamente usado.
- Es el modelo con mayor número de parámetros, lo que implica una serie de problemas, más información adelante.

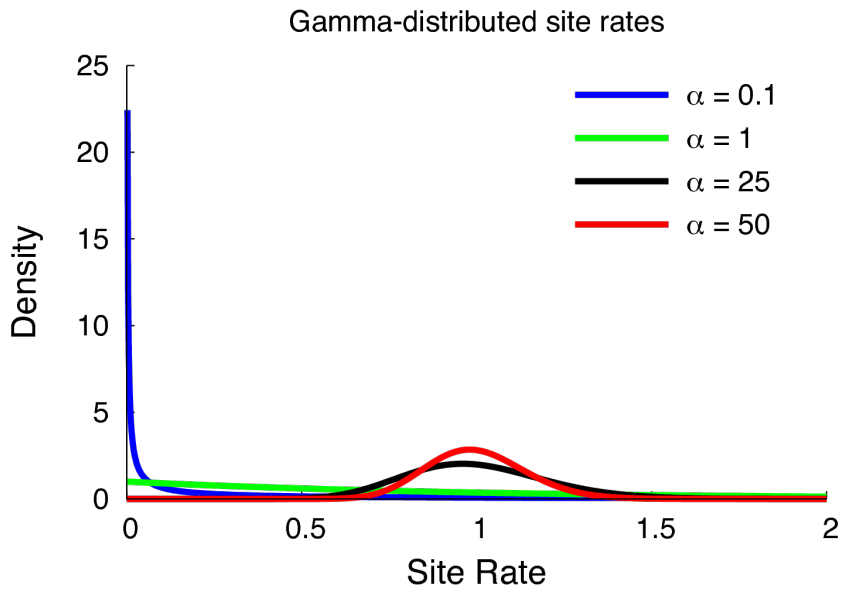


GTR (Tavaré 1986)

Dos adiciones a esta familia de modelos:

- Sitios invariables: Una fracción de las secuencias no evolucionan.
- Distribución gamma de variación en las tasas de substitución

Ambos son ampliamente usados en filogenética.



¿Cómo se selecciona el modelo de evolución?

1. Se estima un árbol filogenético (no tiene que ser necesariamente el real o uno cercano a éste).

¿Cómo se selecciona el modelo de evolución?

1. Se estima un árbol filogenético (no tiene que ser necesariamente el real o uno cercano a éste).
2. Se optimizan los valores de los parámetros para cada uno de los modelos candidatos.

¿Cómo se selecciona el modelo de evolución?

1. Se estima un árbol filogenético (no tiene que ser necesariamente el real o uno cercano a éste).
2. Se optimizan los valores de los parámetros para cada uno de los modelos candidatos.
3. Se estima la verosimilitud (probabilidad) de cada modelo candidato.

¿Cómo se selecciona el modelo de evolución?

1. Se estima un árbol filogenético (no tiene que ser necesariamente el real o uno cercano a éste).
2. Se optimizan los valores de los parámetros para cada uno de los modelos candidatos.
3. Se estima la verosimilitud (probabilidad) de cada modelo candidato.
4. Se aplica un criterio de ganancia de información:
    1. A mayor número de parámetros, la exactitud aumenta (menor sesgo).
    2. A menor número de parámetros, la precisión aumenta (menor varianza).

Tipos de criterio para la selección

- Akaike Information Criterion (AIC).
- Akaike Information Criterion corregido para muestras pequeñas (AICc).
- Bayesian Information Criterion (BIC).

¿Cómo se estima la filogenia?

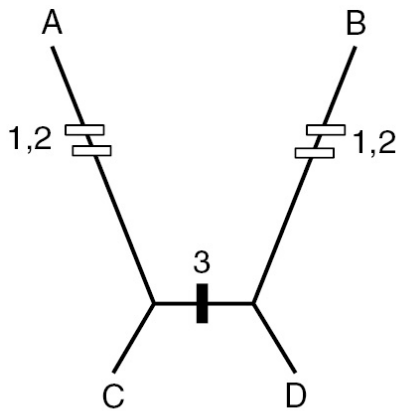
- Se generan árboles de alguna manera (aleatoria o no).
- Los parámetros se optimizan en cada árbol y se estima la probabilidad.
- Se eliminan los árboles de peor probabilidad.
- Con los árboles guardados se pueden generar los árboles de la siguiente iteración.
- Se repite el proceso hasta que el algoritmo converga: no se aumenta la probabilidad del mejor árbol.

Principales programas de estimación filogenética

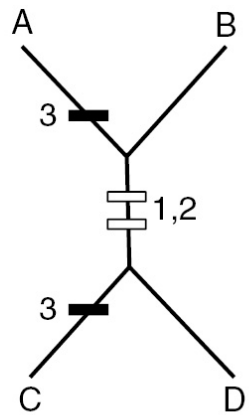
- RAxML.
- PhyML.
- IQTREE.
- Garli.
- PAUP*.
- FastTree

¿Por qué es útil la estimación por ML?

- Escenario: cambios evolutivos no homogéneamente distribuidos en el árbol.
- Parsimonia es consistentemente sesgada en éste escenario.
- Conocido como "atracción de ramas largas" (LBA).
- ML no manifiesta este comportamiento.



"true" tree



most-parsimonious tree

"Atracción de ramas largas" (LBA)

- Sin embargo, LBA es raramente observada en la naturaleza.
- Cuando es observada, ocurre en organismos con cambios drásticos en sus tendencias evolutivas.
- Ejemplo: Evolución de organismos parásitos.

Rafflesia sp.