

Pangenómica y Evolución Microbiana

qBio18, CCG-UNAM, 25 de Julio de 2018

Pablo Vinuesa (vinuesa[AT]ccg.unam.mx)

Programa de Ingeniería Genómica, CCG, UNAM

<http://www.ccg.unam.mx/~vinuesa/>

Temario:

Bloque I: Conceptos de filoinformática para investigación en genómica, ecología y evolución microbiana

- 1) Conceptos básicos de evolución molecular y filogenética
 - 2) Modelos de sustitución nucleotídica e inferencia de filogenias de máxima verosimilitud
 - 3) Práctica: alineamiento múltiple con clustal-omega e inferencia de filogenias con PhyML

Bloque II: Estructura y evolución de genomas microbianos

- 1) Introducción a la filogenómica y pan-genómica microbiana
 - 2) Práctica de pan-genómica - uso del paquete GET_HOMOLOGUES
 - 3) Práctica de filogenómica – uso del paquete GET_PHYLOMARKERS

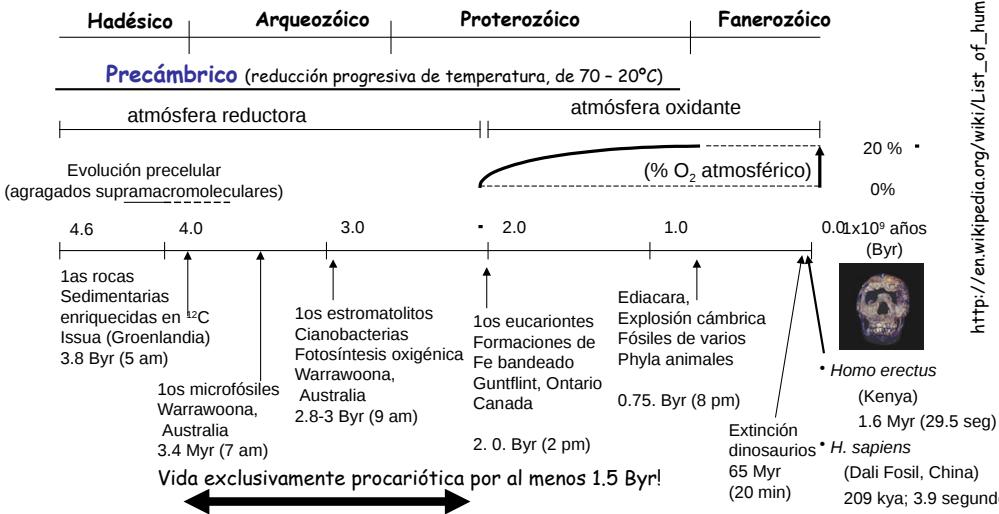
Bloque III: Estudios de caso en genómica evolutiva de microbios

- 1) Presentación del Dr. Pablo Vinuesa, CCG-UNAM
 - 2) Presentación del Dr. Víctor González, CCG-UNAM

<https://github.com/vinuesa/qBio18>

Evolución orgánica - la dimensión temporal

Historia de la tierra y de la vida



2 07/24/2018 21:01 Bloque I:
Conceptos de filo-informática para investigación
en genómica, ecología y evolución microbiana

Pablo Vinuesa, Centro de Ciencias Genómicas - UNAM

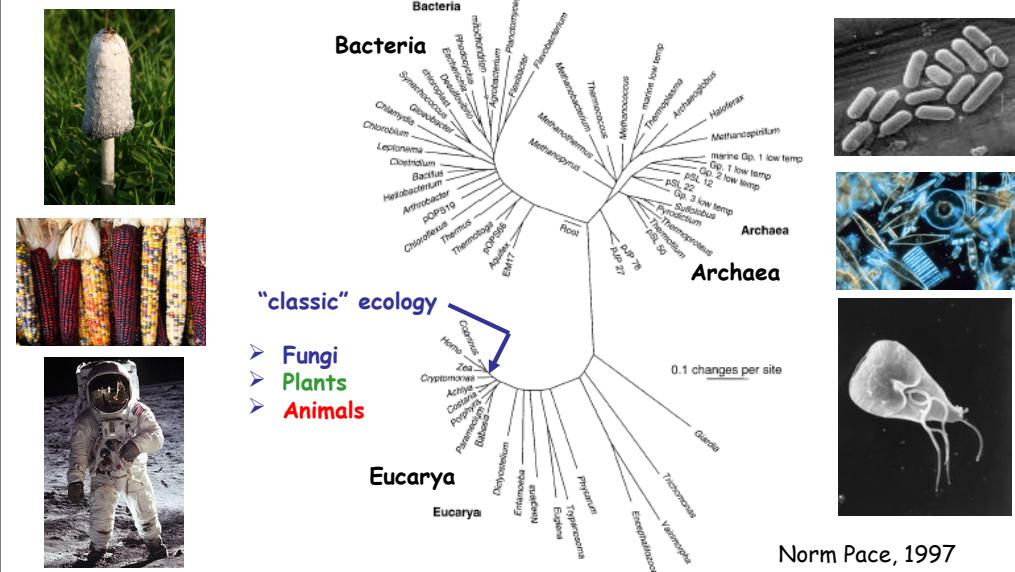
vinuesa@ccg.unam.mx

<http://www.ccq.unam.mx/~vinuesa/>



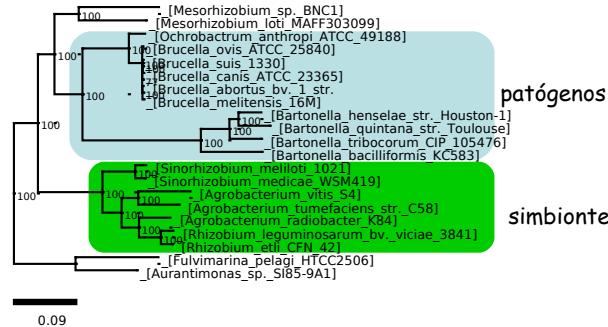
Biodiversity = microbial diversity

- the SSU rDNA view of diversity on Earth



La relación entre filogenética y evolución molecular:

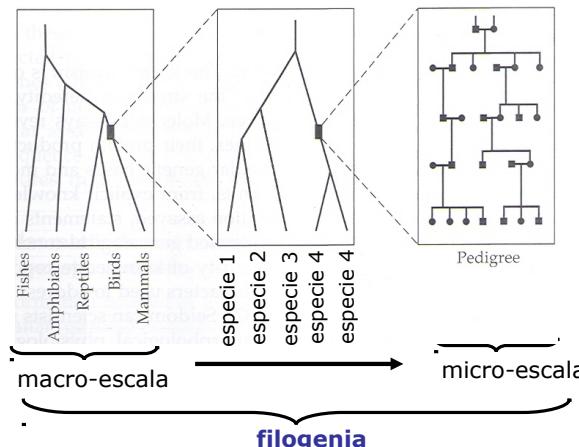
- La **filogenética** tiene por objetivo el **trazar la relación ancestro descendiente de los organismos** (árbol filogenético) a diferentes niveles taxonómicos, incluyendo el árbol universal, haciendo una reconstrucción de esta relación en base a diversos **caracteres homólogos**, tanto **morfológicos** como **moleculares**. Las hipótesis filogenéticas resultantes son la base para hacer **predicciones** (inferencias) sobre propiedades biológicas de los grupos revelados por la filogenia mediante el mapeo de caracteres sobre la topología (hipótesis evolutiva). También proveen el contexto comparativo para poder inferir patrones de **evolución molecular**.



El concepto de filogenia y homología: definiciones básicas

"The stream of heredity makes phylogeny; in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream".

G.G. Simpson (1945)



Filogenia: historia evolutiva del flujo hereditario a distintos niveles evolutivos/temporales, desde la genealogía de genes en poblaciones (micro-escala; dominio de la genética de poblaciones) hasta el árbol universal (macro-escala)

¿Por qué estudiar filogenética y evolución molecular?

Corolario I:

"Nothing in biology makes sense except in the light of evolution"

- Theodosius Dobzhanski, 1973
(*The American Biology Teacher* 35:125)

Corolario II:

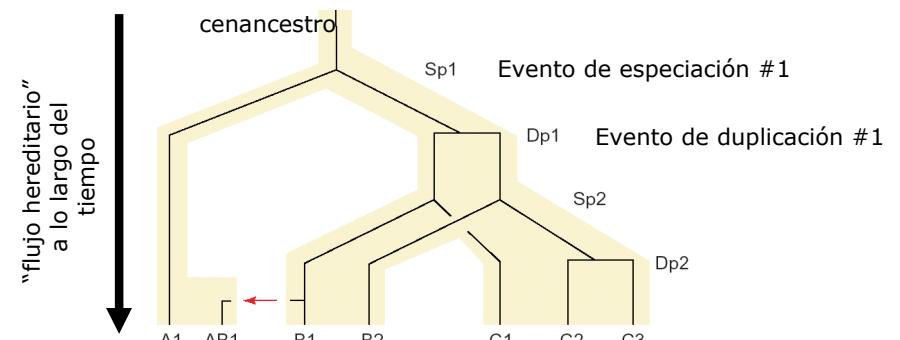
"Nothing in evolutionary biology makes sense except in the light of a phylogeny"

- Jeff Palmer, Douglas Soltis, Mark Chase, 2004
(*American J. Botany* 91: 1437-1445)



La filogenia se infiere a partir de caracteres homólogos

Subtipos de homología: ortología, paralogía y xenología



ortología: relación entre secuencias en la que la divergencia acontece tras un evento de especiación. El ancestro común es el cenancestro. La filogenia recuperada de estas secuencias refleja la filogenia de las especies.

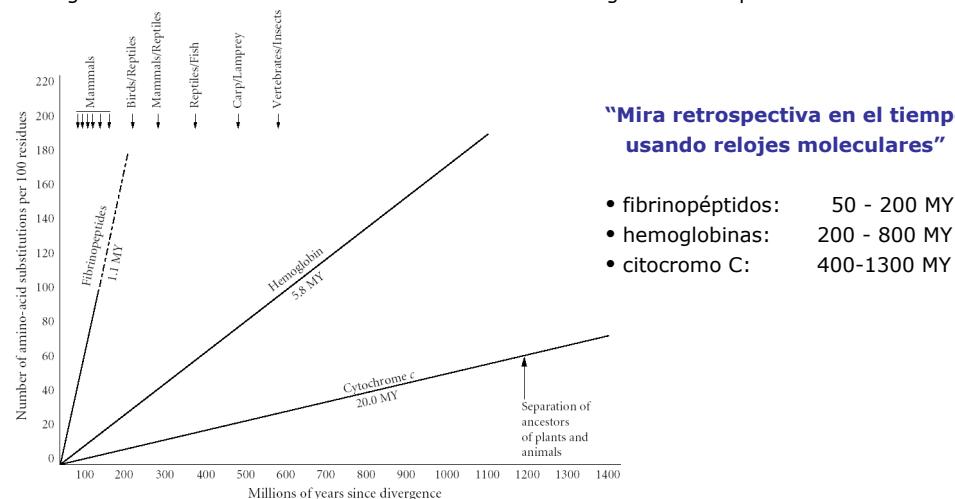
paralogía: condición evolutiva en la que la divergencia observada acontece tras un evento de duplicación génica. La mezcla de ortólogos y parálogos en un mismo análisis filogenético recupera la filogenia correcta de los genes pero no necesariamente la de los organismos o taxa.

xenología: relación entre secuencias dada por un evento de transferencia horizontal entre linajes. Distorsiona fuertemente la filogenia de las especies.

Selección de marcadores adecuados para hacer inferencias evolutivas a distintos niveles de profundidad filogenética

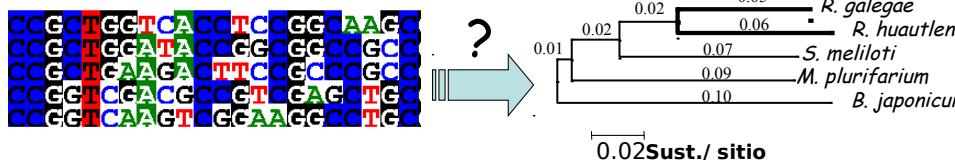
Restricciones funcionales vs. tasas de sustitución:

- Las **tasas de evolución** también varían mucho **entre los genes** de un organismo y del mismo gen en linajes distintos :
 - fibrinopéptidos evolucionan una tasa x900 > a la de ubiquitina y x20 > citocromo C
 - genes de HIV evolucionan a $\times 10^6$ veces la tasa de un gen humano promedio!



Inferencia Filogenética - introducción

- La inferencia de relaciones filogenéticas a partir de secuencias moleculares requiere de la selección de uno de los muchos métodos disponibles
- Con frecuencia la inferencia filogenética es considerada como una "caja negra" en la que "entran las secuencias y salen los árboles"



Objetivos de esta presentación son:

- desarrollar un marco conceptual para entender los fundamentos teóricos (filosóficos) que distinguen a los distintos métodos de inferencia (clasificación de métodos)
- presentar el uso de **modelos y suposiciones** en filogenética
- manejo empírico de diversos paquetes de software para inferencia filogenética bajo diversos criterios

Protocolo básico para un análisis filogenético

de secuencias moleculares

Colección de secuencias homólogas

- BLAST y FASTA

Alineamiento múltiple de secuencias

- Clustal, muscle, T-Coffee ...

Análisis evolutivo del alineamiento y selección del modelo de sustitución más ajustado

- tests de saturación, modeltest, ...

Estima filogenética

- NJ, ME, MP, ML, Bayes ...

Pruebas de confiabilidad de la topología inferida

- proporciones de bootstrap
probabilidad posterior ...

Interpretación evolutiva y aplicación de las filogenias

Inferencia filogenética molecular - clasificación de métodos

- Podemos clasificar a los métodos de reconstrucción filogenética en base al tipo de datos que emplean (**caracteres discretos vs. distancias**) y si usan un **método algorítmico o un criterio de optimización** para encontrar la topología

Tipo de datos

Método de reconstrucción	distancias	caracteres discretos
algoritmo de agrupamiento	UPGMA Neighbour joining	X
criterio de optimización	Evolución mínima	Máxima parsimonia Máxima verosimilitud

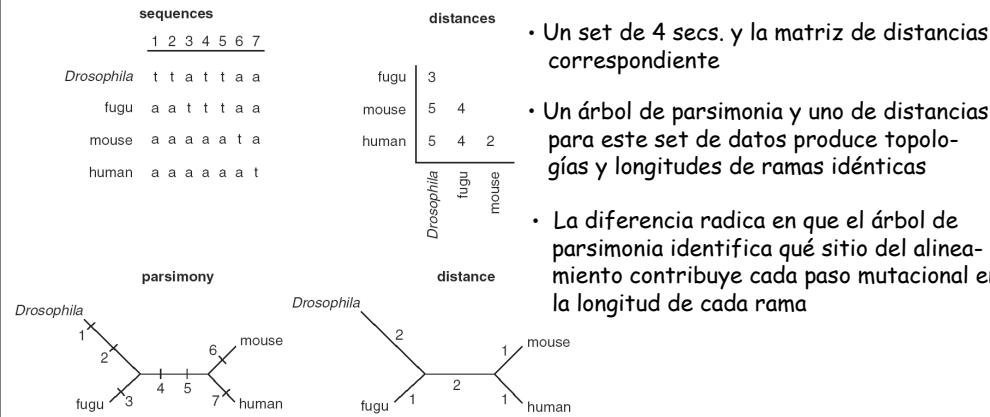
- **Métodos que usan m** modelos explícitos de sustitución
 - métodos de distancia
 - m. de máx. verosimil.
 - m. bayesianos

- **Métodos que no usan** modelos explícitos de sustitución:
 - parsimonia

Métodos de reconstrucción filogenética - una clasificación

I.- Tipos de datos: distancias vs. caracteres discretos

- Los métodos de distancia requieren la transformación de los alineamientos de secuencias en una matriz de distancias genéticas en base al modelo evolutivo seleccionado, la cual es usada por el método algorítmico de reconstrucción para calcular el árbol (UPGMA y NJ)
- Los métodos discretos (MP, ML, Bayesianos) consideran cada sitio del alineamiento (o una función probabilística para cada sitio) directamente



Modelos de evolución de secuencias -introducción

- Para el análisis filogenético de secuencias alineadas virtualmente todos los métodos describen la evolución de las secuencias usando un modelo que consta de dos componentes:

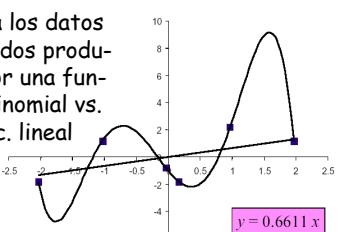
1. un árbol filogenético
2. una descripción de las probabilidades con las que se dan las sustituciones de aa o nts a lo largo de las ramas del árbol

¿Porqué necesitamos modelos y para qué sirven?

- Los modelos nos sirven para interpolar adecuadamente entre nuestras observaciones con el fin de poder hacer predicciones inteligentes sobre observaciones futuras

$$y = -1.5972x^5 + 23.167x^4 - 126.18x^3 + 319.17x^2 - 369.22x + 155.67$$

ajuste a los datos observados producidos por una función polinomial vs. una func. lineal



- añadir parámetros a un modelo generalmente mejora su ajuste a los datos observados
- modelos infra-parametrizados conducen a un pobre ajuste a los datos observados
- modelos supra-parametrizados conducen a una pobre predicción de eventos futuros
- existen métodos estadísticos para seleccionar modelos ajustados a cada set de datos

Modelos de evolución de secuencias -introducción

- Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales
- 1.- La reconstrucción o estima filogenética es un problema de inferencia estadística, y como tal requiere un modelo de sustitución de resíduos (aa o nt), es decir, un modelo de evolución molecular de las secuencias. Todos los modelos, por no ser más que aproximaciones de los procesos naturales, hacen una serie de suposiciones (simplificaciones)
- 2.- Los modelos de evolución de secs. son usados en filogenética para describir las probabilidades con las que se dan los distintos eventos de sustitución entre aa o nt, con el fin de corregir o compensar las sustituciones no observadas a lo largo de la filogenia
- 3.- Mientras que los métodos de MP asumen un modelo implícito de evolución (número mínimo de sustituciones a lo largo de la filogenia), los métodos de distancia (UPGMA, NJ), los de MV y Bayesianos requieren de un modelo explícito de evolución
- 4.- Los métodos de distancia estiman finalmente un sólo parámetro (no. sust./sitio) dado el modelo y el valor de los parámetros del mismo; en cambio, los métodos de ML y Bayesianos pueden estimar el valor de cada uno de los parámetros del modelo explicitado, dada una topología y la matriz de datos (alineamiento)

Modelos de evolución de secuencias -introducción

- Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales
- Existen dos aproximaciones para construir modelos de evolución de secuencias.
 1. construcción de modelos empíricos basados en propiedades del proceso de sustitución calculadas a partir de comparaciones de un gran número de secuencias. Los modelos empíricos resultan en valores fijos de los parámetros, los cuales son estimados sólo una vez, suponiéndose que son adecuados para el análisis de otros sets de datos. Esto los hace fácil de usar e implementar en términos computacionales, pero su utilidad real para cada caso particular ha de ser evaluada críticamente. Es la aproximación utilizada para modelar el proceso de sustitución en proteínas (BLOSUM, PAM ...)
 2. construcción de modelos paramétricos basado en el modelaje de propiedades químicas o genéticas del aas y nts. Los modelos paramétricos tienen la ventaja de que los valores de los parámetros pueden ser derivados de cada set de datos al hacer un análisis de los mismos usando métodos de ML o By, por tanto ajustándolos a cada matriz de datos particular

Modelos de evolución de secuencias -DNA

• Modelos de sustitución de nucleótidos

- El modelaje de la evolución a nivel del DNA se ha concentrado en la aproximación paramétrica. Se manejan **tres tipos principales de parámetros** en estos modelos:

1. parámetros de **frecuencia**

2. parámetros de **tasas de intercambio**

3. parámetros de **heterogeneidad de tasas de sustitución entre sitios**

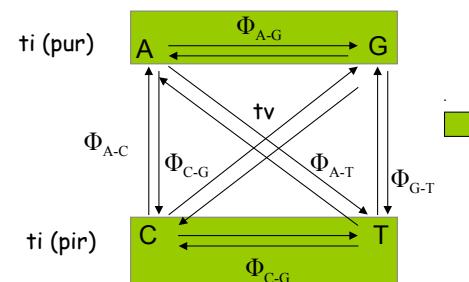
Modelos de evolución de sustitución de nucleótidos -modelos paramétricos

- los diversos modelos evolutivos se distinguen por su grado de parametrización

I. Frecuencias de nt : $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$ ó $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

- modelos de = frecuencia: JC69; K2P, K3P ...
- modelos de \neq frecuencia: F81, HKY85, TrN93, GTR ...

II. Tasas de sustitución transicionales/transversionales

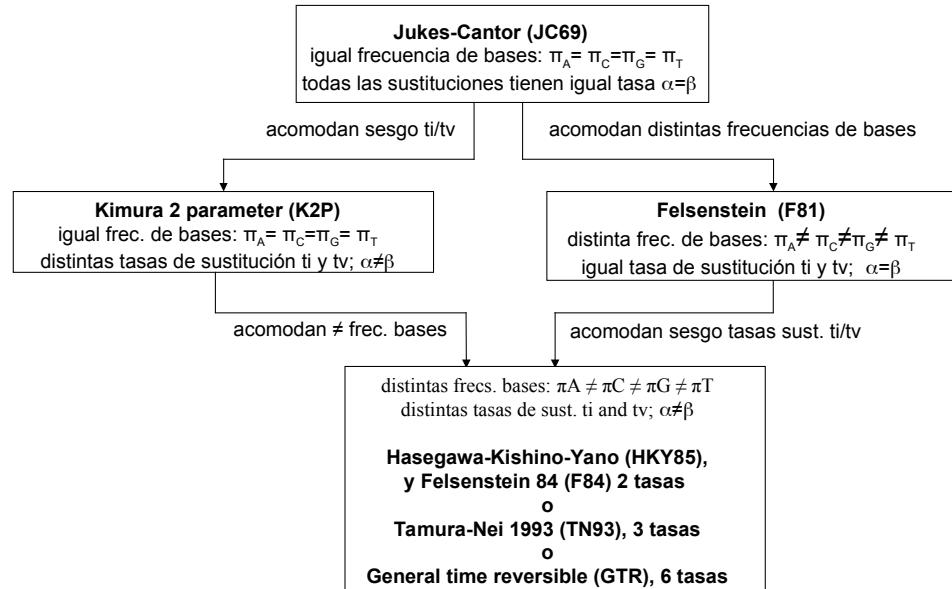


• Existen 4 tipos de sustituciones ti y 8 tv ; cuando $ti/tv \neq 0.5$ existe un sesgo en sustituciones ti (o tv) en el set de datos. ti generalmente $\gg 1$

- los modelos evolutivos se diferencian también en la cantidad de parámetros que utilizan para acomodar diversas tasas de sustitución:

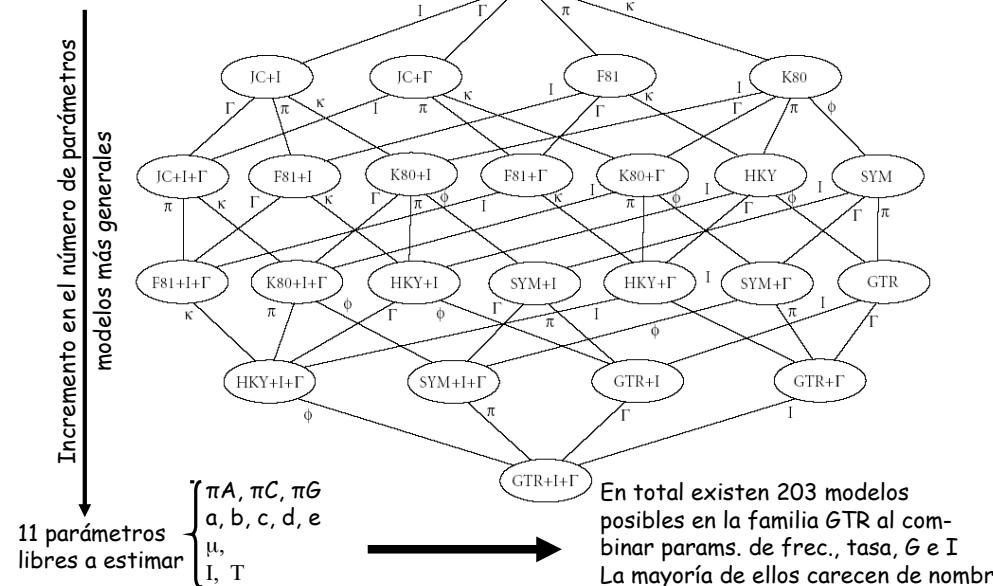
tasas	modelo
1	JC69 ($ti=tv$)
2	K2P ($ti \neq tv$)
3	TrN ó K3P (2 ti , 1 tv)
6	GTR (cada sust. su tasa)

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV



Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

1 parámetro (α)



Comparación empírica de modelos sust. de DNA

- Comparación de los modelos de **JC69** y **K2P** en su capacidad de corregir distancias observadas (p) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p \right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2} \ln \left(\frac{1}{1-2P-Q} \right) + \frac{1}{4} \ln \left(\frac{1}{1-2Q} \right)$$

- Escenario I:

- sean 2 secuencias de long. = 200 nt, que difieren en 20 ti y 4 tv

por lo tanto $L = 200$, $P = 20/200 = 0.1$ y $Q = 4/200 = 0.02$

$$p = 24/200 = 0.12$$

$d_{JC69} \approx 0.13$ (sust./sitio)

$$\text{no. de sust. esperadas} = 0.13 \times 200 \approx 26$$

$d_{K2P} \approx 0.13$ (sust./sitio)

$$\text{no. de sust. esperadas} = 0.13 \times 200 \approx 26$$

Comparación empírica de modelos sust. de DNA

- Comparación de los modelos de **JC69** y **K2P** en su capacidad de corregir distancias observadas (p) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p \right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2} \ln \left(\frac{1}{1-2P-Q} \right) + \frac{1}{4} \ln \left(\frac{1}{1-2Q} \right)$$

- Escenario II:

- sean 2 secuencias de long. = 200 nt, que difieren en 50 ti y 16 tv

por lo tanto $L = 200$, $P = 50/200 = 0.25$ y $Q = 16/200 = 0.08$

$$p = 66/200 = 0.33$$

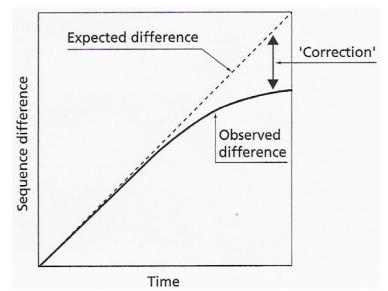
$d_{JC69} \approx 0.43$ (sust./sitio)

$$\text{no. de sust. esperadas} = 0.43 \times 200 \approx 86$$

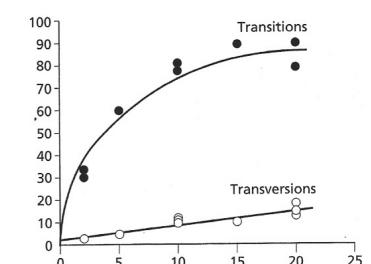
$d_{K2P} \approx 0.48$ (sust./sitio)

$$\text{no. de sust. esperadas} = 0.48 \times 200 \approx 96$$

Modelos de evolución de secuencias



- El objetivo de los modelos de sustitución es el de compensar para los eventos homoplásicos de múltiples sustituciones, y así obtener estimas de distancias evolutivas corregidas



- El número de ti es generalmente $>$ que el de tv , fenómeno que se acentúa cuanto mayor es la divergencia entre las secuencias a comparar. De ahí que en nuestro ejemplo las diferencias entre los escenarios I y II sólo se hicieron notar en el caso en el que la divergencia entre las secuencias era mayor (escenario II)

Inferencia Filogenética y Evolución Molecular - Máxima verosimilitud

Método de agrupamiento	Tipo de datos
distancias	caracteres discretos
criterio de optimización	
UPGMA	
Neighbour joining	
Evolución mínima	Máxima parsimonia Máxima verosimilitud

Criterios de optimización II - Máxima verosimilitud (ML) y selección de modelos de sustitución

1. El criterio de optimización de máxima verosimilitud en filogenética
2. ML y estima de parámetros del modelo de sustitución
3. ML y contraste de hipótesis evolutivas (selección de modelos (LRT, AIC)

Maxima verosimilitud y estima de parámetros de modelos de sustitución

- La inferencia filogenética bajo el criterio de máxima verosimilitud se basa en el uso de una cantidad llamada **log-likelihood** para evaluar topologías alternativas con el fin de encontrar aquella que **maximiza este valor**.
- El **log-likelihood** es el \ln de la verosimilitud, que es igual a la probabilidad de los datos observados dadas una topología particular (τ), set de longitudes de rama (v) y modelo de sustitución (ϕ).
- Nótese que **la verosimilitud no representa la probabilidad de que un árbol sea correcto**; ésta viene determinada por la **probabilidad posterior** de la estadística bayesiana.
- Harlar de la "verosimilitud de un conjunto de datos" no es correcto ya que la verosimilitud es un función de los parámetros de un modelo estadístico, y no de los datos (D). **Los datos son constantes siendo el modelo lo que es variable al calcular verosimilitudes**. Se puede por lo tanto hablar de verosimilitudes como funciones de modelos o hipótesis (H). La verosimilitud de una hipótesis dado un set de datos es igual a la probabilidad condicional de los datos dada una hipótesis.

Formalmente: $L(H|D) = \Pr(D|H) = \Pr(D|\tau v \phi)$

Metodos de reconstrucción filogenética - Máxima Verosimilitud

Máxima verosimilitud: dadas dos topologías, la que hace los datos observados más probables ("menos sorprendentes") es la preferida

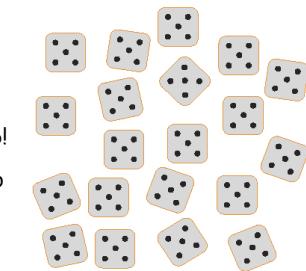
El **método de máxima verosimilitud (ML)** considera cada sitio variable del alineamiento (incluidos singletones). Bajo el criterio de ML se busca la topología que hace más verosímil el patrón de sustituciones de un alineamiento dado un modelo evolutivo explícito!

Así, para un set de datos D y una hipótesis evolutiva (topología) H , la verosimilitud de dichos datos viene dado por la expresión:

$$L_D = \Pr(D|H)$$
 que es la probabilidad de obtener D dada H (una **probabilidad condicional**) !

Por tanto **la topología que hace nuestros datos el resultado evolutivo más probable corresponde a la estima de máxima verosimilitud de la filogenia** (likelihood score ó valor de verosimilitud).

- la probabilidad está relacionada con la "sorpresividad" de los datos
- Estaríamos sorprendidos de obtener este resultado, dada su bajísima probabilidad $(1/6)^{20}$ ó 1 en 3,656,158, 440,062,976!
- Pero la probabilidad depende del modelo probabilístico asumido
- En filogenética, las distintas topologías representan a los distintos modelos, y se selecciona aquel modelo que nos hace sorprendernos menos de los datos que hemos coleccionado



Maxima verosimilitud y estima de parámetros de modelos de sustitución

$$L(H|D) = \Pr(D|H) = \Pr(D|\tau v \phi)$$

- Lo mejor es pensar en los **árboles como modelos**. La verosimilitud de una topología particular (τ) será la probabilidad de los datos dada esa topología. Cada topología tiene como parámetros las longitudes de rama (v), y la verosimilitud de un modelo (ϕ) cambia según varíen los valores de los parámetros de longitud de rama
- Por lo tanto se puede concebir la filogenética bajo el criterio de máxima verosimilitud como un **problema de selección de modelos**. Se trata de encontrar las estimas de los valores de cada parámetro del modelo y luego comparar las verosimilitudes de los distintos modelos, escogiendo el mejor (topología) en base a su verosimilitud
- La topología que hace de nuestros datos el resultado evolutivo más probable (dado un modelo de sust.) es la estima de máxima verosimilitud de nuestra filogenia. Por tanto, al contrario que bajo los criterios de optimización de MP, LS o ME, **bajo ML se trata de seleccionar modelos y parámetros que maximicen la función de optimización**.

Maxima verosimilitud y estima de parámetros de modelos de sustitución

- Cálculo del valor de máxima verosimilitud para una sola secuencia o árbol trivial con un solo nodo

primeros 25 nt del gen *ropB* de *Bradyrhizobium japonicum* USDA110

ATGGCGCAGCAGACATTCAACGGTC

$$L = \pi_A \pi_T \pi_G \pi_C \pi_G \pi_C \pi_A \pi_G \pi_C \pi_A \pi_G \pi_A \pi_C \pi_A \pi_T \pi_C \pi_A \pi_C \pi_G \pi_G \pi_T \pi_C$$

$$= \pi_A^{nA} \pi_C^{nC} \pi_G^{nG} \pi_T^{nT} = \pi_A^6 \pi_C^8 \pi_G^7 \pi_T^4$$

$$\pi_A = 0.24$$

$$\pi_C = 0.32$$

$$\pi_G = 0.28$$

$$\pi_T = 0.16$$

$$\ln L = 6 \ln (\pi_A) + 8 \ln (\pi_C) + 7 \ln (\pi_G) + 4 \ln (\pi_T)$$

- A primera vista podemos sospechar que el modelo de F81 se va a ajustar mejor a los datos que el de JC69, ya que las frecuencias de nucleótidos difieren claramente de 0.25, con exceso de Cs y defecto de Ts

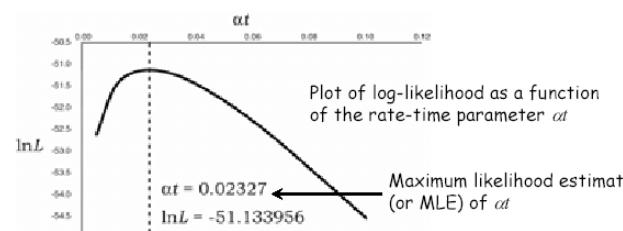
Maxima verosimilitud y estima de parámetros de modelos de sustitución

- Estima del parámetro compuesto αt del modelo JC69 para los primeros 30 nts de la $\gamma\eta$ globina de gorila y orangután

gorilla GAACTCCTTGAGAAATAACTGCACACACTGG
orangutan GCACTCCTTGAGAAATAACTGCACACACTGG

$$L = \left[\left(\frac{1}{4} \right) \left(\frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \right) \right]^{28} \left[\left(\frac{1}{4} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \right]^2$$

- ¿Cómo estimamos el valor de αt ? La estima de máxima verosimilitud se obtiene del análisis de la función de verosimilitud, esencialmente probando diversos valores para el parámetro y determinando cual maximiza la función



$$d_{JC69} = 3\alpha t \\ = 3(0.0237) \\ = 0.0474$$

Maxima verosimilitud y estima de parámetros de modelos de sustitución

- Cálculo del valor de máxima verosimilitud para una sola secuencia o árbol trivial con un solo nodo

primeros 25 nt del gen *ropB* de *Bradyrhizobium japonicum* USDA110

ATGGCGCAGCAGACATTCAACGGTC

- Cálculo de $\ln L$ bajo el modelo de JC69

$$\begin{aligned} \ln L &= 6 \ln (\pi_A) + 8 \ln (\pi_C) + 7 \ln (\pi_G) + 4 \ln (\pi_T) \\ &= 6 \ln (0.25) + 8 \ln (0.25) + 7 \ln (0.25) + 4 \ln (0.25) = -29.1 \end{aligned}$$

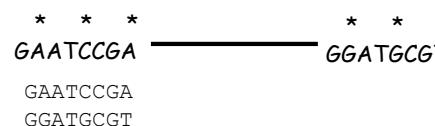
- Cálculo de $\ln L$ bajo el modelo de F81

$$\begin{aligned} \ln L &= 6 \ln (\pi_A) + 8 \ln (\pi_C) + 7 \ln (\pi_G) + 4 \ln (\pi_T) \\ &= 6 \ln (0.24) + 8 \ln (0.32) + 7 \ln (0.28) + 4 \ln (0.16) = -26.6 \end{aligned}$$

-Por lo tanto el modelo de F81 se ajusta mejor a los datos ($-26.6 > -29.1$). Esta diferencia será tanto más notoria cuanto más larga sea la secuencia.

Maxima verosimilitud y estima de parámetros de modelos de sustitución

- Esquema del procedimiento del cálculo del valor de verosimilitud de un árbol con 4 OTUs

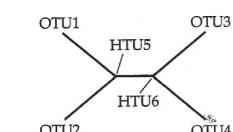


$$L = L_1 L_2 \dots L_8 = [1/16 (1 + 3e^{-4\alpha t})]^5 [1/16 (1 - e^{-4\alpha t})]^3$$

• En un "árbol" con sólo 2 OTUs no tenemos ningún nodo interior o ancestral. El cómputo lo realizamos directamente sobre los datos observados

- La complicación adicional que encontramos para el cálculo de verosimilitudes de árboles con > 3 OTUs radica esencialmente en que tenemos ahora nodos interiores para los que carecemos de observaciones. Se trata de unidades taxonómicas hipotéticas HTUs. En este caso, para calcular la verosimilitud del árbol tenemos que considerar cada posible estado de carácter para cada nodo interior y para cada topología !!!.

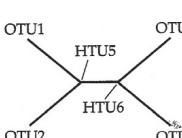
OTU1	1	2	3	4	5	6	7	8	9	...n
OTU2	A	G	C	C	C	T	T	C	A	...
OTU3	A	G	A	T	A	T	C	C	A	...
OTU4	A	G	A	G	G	T	C	C	T	...



Maxima verosimilitud y estima de parámetros de modelos de sustitución

- Esquema del procedimiento del cálculo del valor de verosimilitud de un árbol con 4 OTUs

1	2	3	4	5	6	7	8	9	...n	
OTU1	A	A	G	A	C	T	T	C	A	...N
OTU2	A	G	C	C	C	T	T	C	T	...N
OTU3	A	G	A	T	A	T	C	C	A	...N
OTU4	A	G	A	G	G	T	C	C	T	...N



$$L_{(5)} = \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} A & A \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} A & C \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} A & T \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} A & G \\ C & G \end{array} \right)$$

$$+ \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} C & A \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} C & C \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} C & T \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} C & G \\ C & G \end{array} \right)$$

$$+ \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} T & A \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} T & C \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} T & T \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} T & G \\ C & G \end{array} \right)$$

$$+ \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} G & A \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} G & C \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} G & T \\ C & G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ C \end{array} \middle| \begin{array}{cc} G & G \\ C & G \end{array} \right)$$

$$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \dots \times L_{(n)} = \prod_{i=1}^n L_{(i)}$$

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + \dots + L_{(n)} = \sum_{i=1}^n \ln L_{(i)}$$

• La **verosimilitud para cada sitio** representa la suma sobre todas las posibles asignaciones de estados de carácter en todas las ramas interiores de un árbol. La **verosimilitud total** es el producto de las veros. por sitio.

Maxima verosimilitud y estima de parámetros de modelos de sustitución

3. Prueba de razón de verosimilitudes (LRT)

- Una manera natural y muy usada de comparar el ajuste relativo de dos modelos alternativos a una matriz de datos es contrastar las verosimilitudes resultantes mediante la prueba de razones de verosimilitud (RV) ó likelihood ratio test (LRT):

$$\Delta = 2(\log_e L_1 - \log_e L_0)$$

donde L_1 es el valor de ML global para la hipótesis alternativa (modelo más rico en parámetros) y L_0 es el valor de ML global para la hipótesis nula (el modelo más simple).

$\Delta > 0$ siempre, ya que los parámetros adicionales van a dar una mejor explicación de la variación estocástica en los datos que el modelo más sencillo.

• Cuando los modelos a comparar están anidados (L_0 es un caso especial de L_1) el estadístico Δ sigue aproximadamente una distribución X^2 con q grados de libertad, donde

q = diferencia entre el no. de parámetros libres entre L_1 y L_0 .

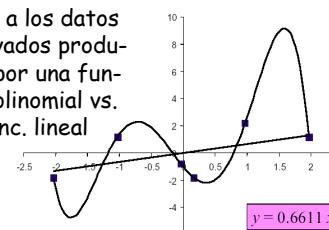
Maxima verosimilitud y estima de parámetros de modelos de sustitución

2. Selección de modelos de sustitución de secuencias de DNA

- En términos generales modelos complejos se ajustan a los datos mejor que los simples. Idealmente se ha de seleccionar un modelo lo suficientemente complejo (rico en parámetros) como para describir adecuadamente las características más notables del patrón de sust. del set de datos, pero no sobreparametrizado para evitar colinealidad de parámetros (redundancia), tiempos excesivamente largos de cómputo y estimas poco precisas de los parámetros por excesiva varianza.

$$y = -1.5972 x^5 + 23.167 x^4 - 126.18 x^3 + 319.17 x^2 - 369.22 x + 155.67$$

ajuste a los datos observados producidos por una función polinomial vs. una func. lineal



- añadir parámetros a un modelo generalmente mejora su ajuste a los datos observados
- modelos infra-parametrizados conducen a un pobre ajuste a los datos observados
- modelos supra-parametrizados conducen a una pobre predicción de eventos futuros
- existen métodos estadísticos para seleccionar modelos ajustados a cada set de datos

Maxima verosimilitud y estima de parámetros de modelos de sustitución

3. Prueba de razón de verosimilitudes (LRT)

- El LRT es por tanto una prueba estadística para cuantificar la bondad relativa de ajuste entre dos modelos anidados. Veamos un ejemplo. Vamos seleccionar entre los modelos JC69, F81, HKY85 y TrN93 para el set de datos de mtDNA-primates.nex, considerando sólo las regiones codificadoras y eliminando Lemur_catta, Tarsius_syrichta y Saimiri_scireus y usando un árbol NJ sobre el cual estimar parámetros

Modelo	$-\ln L$
JC69	3585.54820
F81	3508.04085
HKY85	3233.34395
TrN93	3232.29439

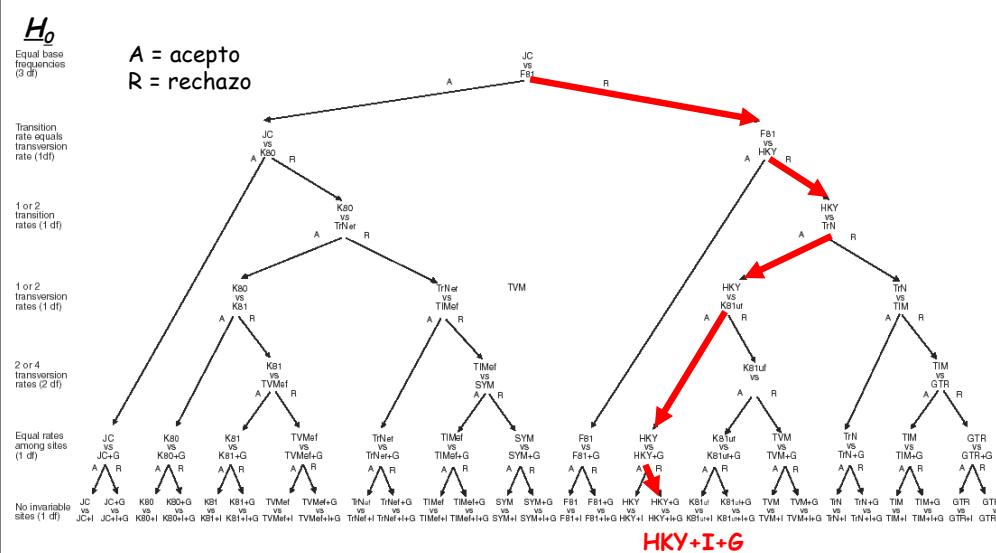
- ¿Qué podemos concluir de estos valores de $-\ln L$ en cuanto a la importancia relativa de los parámetros considerados por estos modelos en cuanto al nivel de ajuste a los datos que alcanzan?

Maxima verosimilitud y estima de parámetros de modelos de sustitución**3. Prueba de razón de verosimilitudes (LRT)**

<u>Modelo</u>	<u>-lnL</u>	<u>H_0 a rechazar (o hipótesis anidadas a evaluar)</u>	
JC69	3585.54820		
F81	3508.04085	1. igual freq. de bases	
HKY85	3233.34395	2. $T_i = T_v$	
TrN93	3232.29439	3. tasas de T_i iguales	
		...	
modelos	diff. GL = q	X²	P
JC-F81	3 - 0 = 3	155	0
JC-HKY85	4 - 0 = 4	704.4	0
JC-TrN	5 - 0 = 5	706.4	0
F81-HKY85	4 - 3 = 1	549.4	0
F81-TrN	5 - 3 = 2	551.4	0
HKY-TrN	5 - 4 = 1	2.1	0.15

Por lo tanto el modelo seleccionado es el **HKY**

<http://www.fournilab.ch/rpkp/experiments/analysis/chicCalc.html>

Maxima verosimilitud y estima de parámetros de modelos de sustitución**3. Esquema jerárquico de efectuar LRTs partiendo desde el modelo más sencillo (JC69)****Maxima verosimilitud y estima de parámetros de modelos de sustitución****3. Prueba de razón de verosimilitudes (LRT)**

<u>Modelo</u>	<u>-lnL</u>	<u>H_0 a rechazar (o hipótesis anidadas a evaluar)</u>	
HKY85	3233.34395		
HKY85 +G	3145.29031	1. tasa homogénea de sust entre sitios	
HKY85 +I+G	3142.36439	2. no existe proporción de sitios invariantes	

<u>modelos</u>	<u>diff. GL = q</u>	<u>X²</u>	<u>P</u>	
HKY85-vs. +G	1	176	0	Por lo tanto el modelo seleccionado es el HKY+G
HKY85+G vs. I+G	1	5.85	0.015	si tomamos 0.01 como punto de corte, o HKY+I+G si usamos alfa = 0.05.

Maxima verosimilitud y estima de parámetros de modelos de sustitución**3. Selección de modelos usando criterios de información**

- LRT compara pares de modelos anidados. Los criterios de información como el **Akaike information criterion (AIC)** y **Bayesian information criterion (BIC)** comparan simultáneamente todos los modelos en competición y permiten seleccionar modelos aunque no sean anidados.

- Se trata nuevamente de incorporar tanta complejidad (parámetros) al modelo como requieran los datos. La verosimilitud para cada modelo es penalizada en función del número de parámetros: **a mayor cantidad de parámetros mayor penalización**.

Maxima verosimilitud y estima de parámetros de modelos de sustitución

3. Selección de modelos usando criterios de información

- **AIC.** Es un estimador no sesgado del parámetro de contenido de información de Kullback-Leibler, el cual es una medida de la información perdida al usar un modelo para aproximar la realidad. Por tanto, a menor valor de AIC mejor ajuste del modelo a los datos. Al penalizar por cada parámetro adicional, considera tanto la bondad de ajuste como la varianza asociada a la estima de los parámetros.

$$AIC_i = -2 \ln L_i + 2 N_i \quad N_i = \text{no. de parámetros libres en el modelo } i$$

L_i = verosimilitud bajo el modelo i

Maxima verosimilitud y estima de parámetros de modelos de sustitución

3. Selección de modelos usando criterios de información: AIC

- Las ponderaciones o pesos de Akaike (w_i) son los AIC relativos normalizados para cada modelo en competición y pueden ser interpretados como la probabilidad de que un modelo es la mejor abstracción de la realidad dados los datos. Para R modelos candidatos a evaluar:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^R \exp\left(-\frac{1}{2}\Delta_r\right)}$$

- Una aplicación muy útil de los w_i es que la inferencia se puede promediar a partir de los modelos que muestran valores de no w_i triviales. Así, una estima del valor del parámetro a de la distribución gamma promediada a partir de varios modelos se calcularía así:

$$\hat{\bar{\alpha}} = \sum_{i=1}^R w_i \hat{\bar{\alpha}}_i$$

También podemos reconstruir filogenias bajo los distintos modelos con peso significativo y combinar los árboles resultantes acorde a sus pesos de Akaike. Esta estrategia es particularmente útil en un contexto bayesiano.

Maxima verosimilitud y estima de parámetros de modelos de sustitución

3. Selección de modelos usando criterios de información: AIC

- Se pueden usar los estadísticos de diferencias en AIC (Δ_i) y ponderaciones de Akaike para cuantificar el nivel de incertidumbre en la selección del modelo. Las Δ_i son AICs re-escalados con respecto al modelo con el AIC más bajo (minAIC), de modo que $\Delta_i = AIC_i - \text{minAIC}$.
- Las Δ_i son fáciles de interpretar y permiten ordenar los los modelos candidatos. Así, modelos con Δ_i en un rango de 1-2 con respecto al modelo ganador tienen un soporte sustancial y deben de ser considerados como modelos alternativos. Modelos con Δ_i en un rango de 3-7 con respecto al modelo ganador tienen un soporte significativamente inferior, y modelos con $\Delta_i > 10$ carecen de soporte.

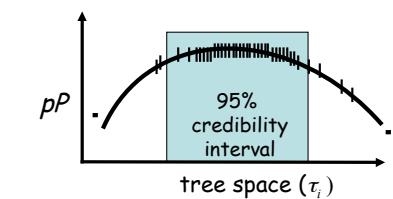
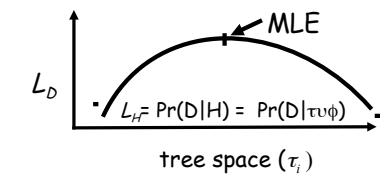
Maxima verosimilitud y estima de parámetros de modelos de sustitución

3. Selección de modelos usando criterios de información: AIC

- Las ponderaciones o pesos de Akaike (w_i) son los AIC relativos normalizados para cada modelo en competición y pueden ser interpretados como la probabilidad de que un modelo es la mejor abstracción de la realidad dados los datos. Para R modelos candidatos a evaluar:

Criterios de optimización: la alteranativa Bayesiana

- Aproximaciones tradicionales (matrices de distancia, ME, ML, MP)
 - la búsqueda tiene por objetivo encontrar la topología óptima (estima puntual)
 - no pueden establecer el soporte relativo de las biparticiones a partir de una única búsqueda
- Aproximación Bayesiana
 - no busca una sola topología óptima sino una población de árboles muestreados en función de su probabilidad posterior (algoritmos MCMC)
 - la muestra de árboles obtenidos en una sola sesión de "búsqueda" es usada para valorar el soporte de cada split en términos de probabilidad posterior



• Prácticas con jmodeltest2 y phylm

1. selección del modelo más ajustado usando [jmodeltest2](#)
2. Estima de una filogenia de máxima verosimilitud usando [phylm3](#)

• Tarea:

1. Revisa los apuntes de esta clase y los tutoriales de uso de jmodeltest y phylm3 desde la línea de comandos.
2. Usa el alineamiento múltiple de codones de secuencias *rpoB* de *Bradyrhizobium* que generaste en la clase pasada para estimar una filogenia de máxima verosimilitud bajo el modelo más ajustado, seleccionado con jmodeltest2.

Tienes que entregar los comando usados y una imagen de la filogenia resultante (generalas con FigTree). Incorpórala al archivo word que subirás junto con los comandos al sitio del curso.

• [Software recomendado para la generación y edición de alineamientos múltiples, inferencia filogenética y visualización de árboles](#)

1.- Alineamientos múltiples y su edición

- BioEdit (sólo Windows)
- jalview y [seaview](#) (multiplataforma)
- [clustalo](#), [clustalw](#)

2.- Paquetes y programas de inferencia filogenética:

- DAMBE y MEGA6 (sólo Windows)
- PAUP* (es el único no libremente disponible en la red)
- [PHYLIP 3.69](#) (multiplataforma)
- [PhyML3](#) (multiplataforma)

3.- Edición y visualización de árboles

- MEGA6 (sólo Windows)
- [FigTree](#) (multiplataforma)

- Una extensa y actualizada lista de programas usados en filogenética la puedes encontrar en el sitio web de Joe Felsenstein
<http://evolution.genetics.washington.edu/phylip/software.html>
- Y en mi sitio web tengo páginas sobre recursos de software para filoinformática
http://www.ccg.unam.mx/~vinuesa/filoinfo_IE11/recursos_bioinfo.html
http://www.ccg.unam.mx/~vinuesa/filoinfo_IE11/recursos_filogenet.html

Slide 52 07/24/2018 21:01 Métodos de búsqueda de árboles

• Pasos lógicos de los métodos filogenéticos basados en criterios de optimización (MP, ML ...)

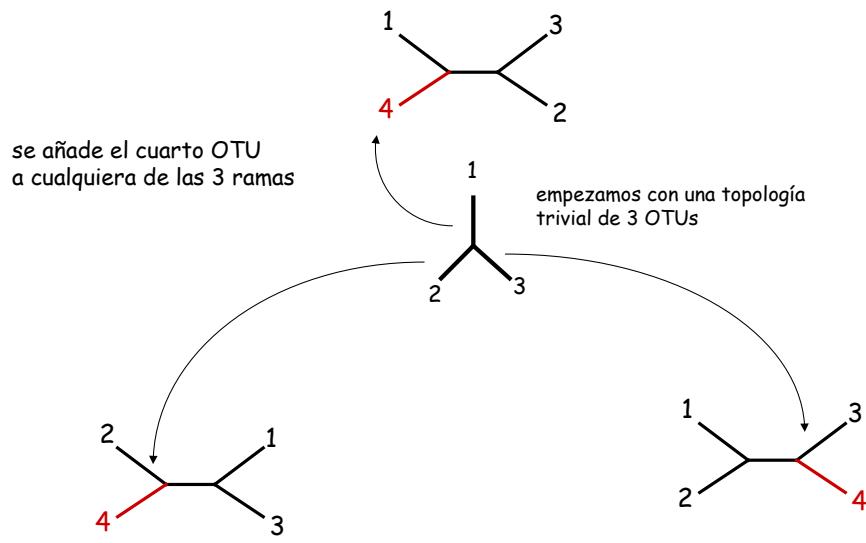
1. definir el criterio de optimización (descrito formalmente en una [función objetiva](#))
2. Construir un árbol de partida que contenga todos los OTUs
3. Emplar [algoritmos de búsqueda](#) que tratan de encontrar árboles mejores bajo el criterio de optimización escogido que el árbol actual o de partida.

1. Criterios de optimización	2. Estrategias de búsqueda
Parsimonia	Enumeración exhaustiva ($n \leq 12$) (exhaustive enumeration)
Máxima verosimilitud	Ramificación y límite ($n \leq 25$) (branch-and-bound)
Evolución Mínima	Decomposición en estrella (star decomposition)
Mínimos cuadrados	Adición secuencial (stepwise addition)
	(Inter-)cambio de rama (branch swapping)

Métodos exactos:
 garantizan encontrar la topología óptima

Métodos heurísticos:
 no garantizan encontrar la topología óptima

Métodos de búsqueda de árboles -enumeración exhaustiva ($n \leq 12$)

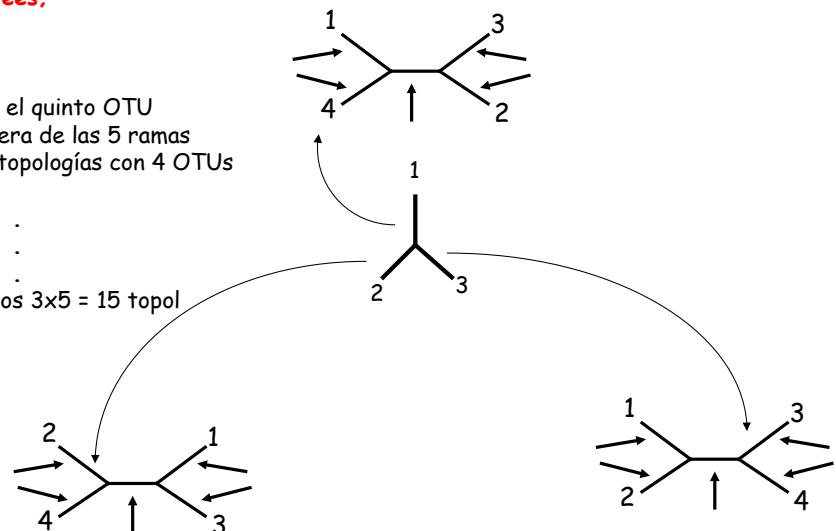


Métodos exactos de búsqueda de árboles -enumeración exhaustiva ($n \leq 12$)

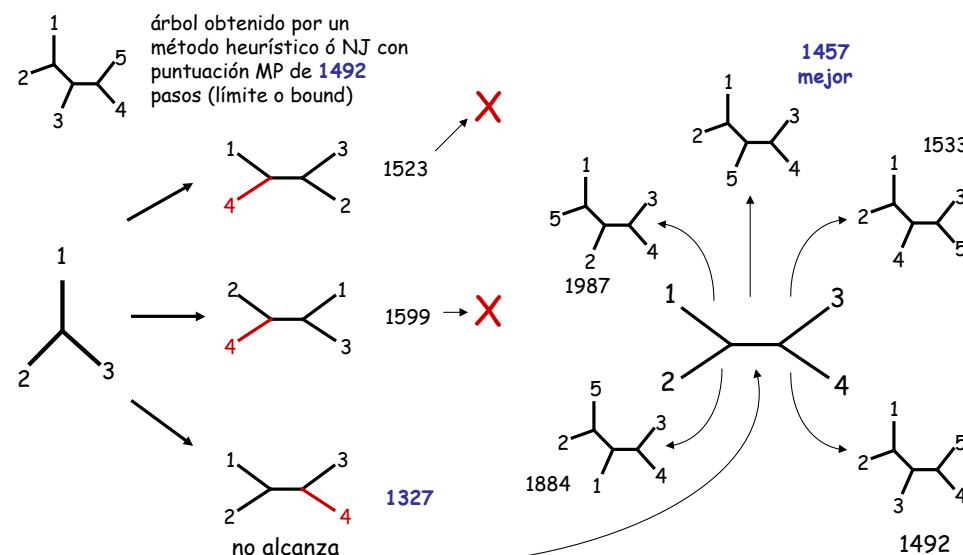
PAUP* command:
`alltrees:`

se añade el quinto OTU a cualquiera de las 5 ramas de las 3 topologías con 4 OTUs

obtenemos $3 \times 5 = 15$ topol



Métodos exactos de búsqueda de árboles - "branch and bound" ($n \leq 25$)



• PAUP* command:
`bandb;`

• Al igual que la búsqueda exhaustiva, garantiza encontrar el árbol óptimo

Métodos de búsqueda de árboles

I.- el problema del número de topologías

El número de topologías posibles incrementa factorialmente con cada nuevo taxon o secuencia que se añade al análisis

$$\text{No. de árboles no enraizados} = (2n-5)!/2^{n-3}(n-3)$$

$$\text{No. de árboles enraizados} = (2n-3)!/2^{n-2}(n-2)$$

Taxa	árboles no enraiz.*	árб. enraiz.
4	3	15
8	10,395	135,135
10	2,027,025	34,459,425
22	3×10^{23}	...
50	3×10^{74}	...

*por ej. para sólo 15 OTUs tenemos 213,458,046,676,875 topologías

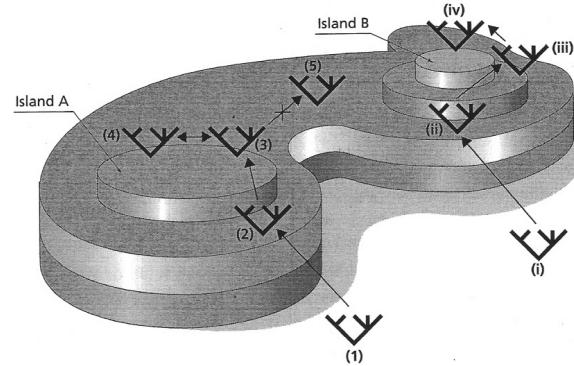
- si pudiésemos evaluar 1×10^6 topol./seg. necesitaríamos 6 años y 9 meses para completar la búsqueda! El no. de Avogadro es $\sim 6 \times 10^{23}$ (átomos/mol). Según la teor. de la relatividad de la estructura del universo de Einstein, existen 10^{80} átomos de H₂ en el universo ...

http://en.wikipedia.org/wiki/Observable_universe

Por tanto se requieren de **estrategias heurísticas de búsqueda** árboles cuando se emplean métodos basados en criterios de optimización y $n > \sim 25$

Métodos heurísticos de búsqueda de árboles - islas de árboles

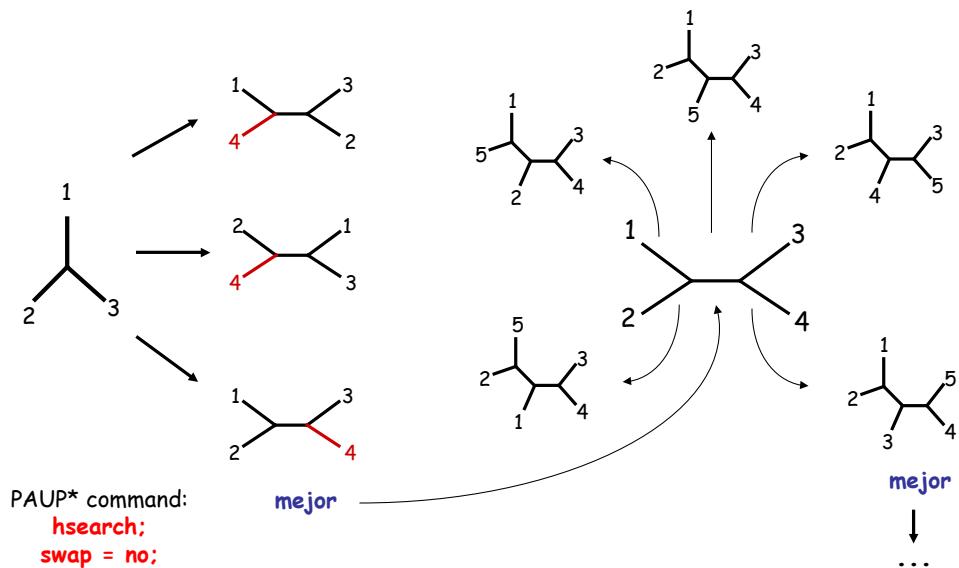
- En la mayor parte de los análisis emplearán métodos heurísticos;
- éstos comienzan con un árbol (aleatorio, NJ o de adición secuencial) para realizar intercambios de ramas (**branch swappig**) sobre esta topología inicial con el propósito de encontrar topologías de mejor puntuación (según la func. de objetividad) que la de partida
- estos métodos heurísticos no garantizan encontrar la topología óptima pero trabajan muy bien cuando se comparan con sets de datos de ≤ 25 secs. analizados mediante B&B



- El espacio de árboles puede visualizarse como un paisaje con colinas de diversas alturas; cada pico representa un máximo local de score o puntuación (**isla de árboles igualmente parsim.**)
- Es recomendable hacer múltiples búsquedas heurísticas, comenzando cada una desde una topología distinta para minimizar el riesgo de obtener un árbol ubicado en una isla topológica subóptima

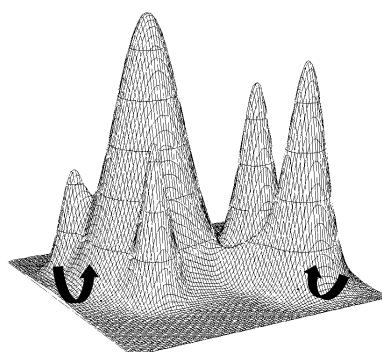
Métodos heurísticos de búsqueda de árboles - adición secuencial (aleatorizada)

Este método se usa con frecuencia para generar distintos "árboles semilla" a partir de los cuales comenzar búsquedas heurísticas, partiendo de "distintos puntos del espacio de árboles"



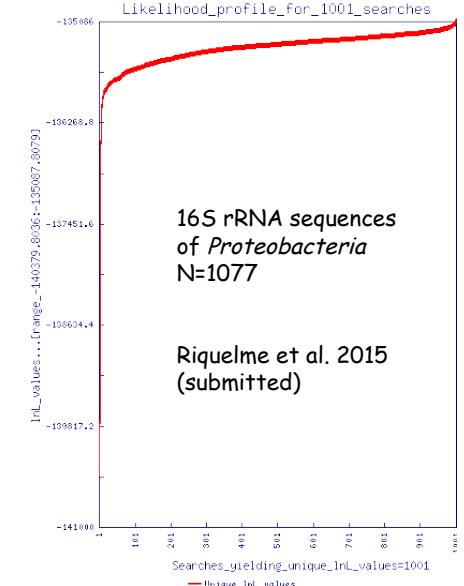
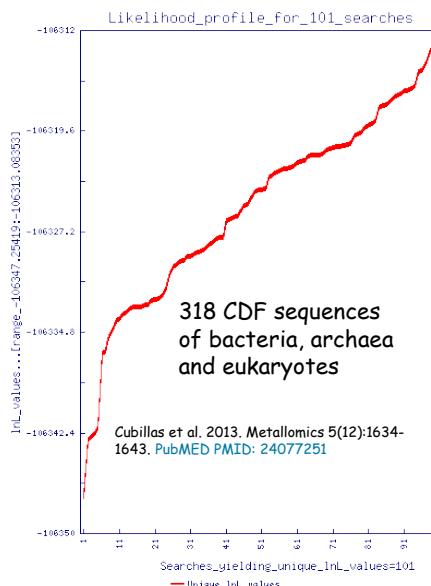
Métodos heurísticos de búsqueda de árboles - adición secuencial (aleatorizada)

- El orden en el que se añaden los OTUs puede cambiar los resultados
- Por ello suele repetirse varias veces, añadiendo OTUs en cada ciclo de manera aleatorizada
- Sirven por lo tanto como **árboles semilla** para iniciar distintas búsquedas heurísticas partiendo de topologías potencialmente diferentes para eficientizar la exploración del espacio de topologías (pero **no adecuados como hipótesis filogenética en sí mismos**)



Métodos heurísticos de búsqueda de árboles

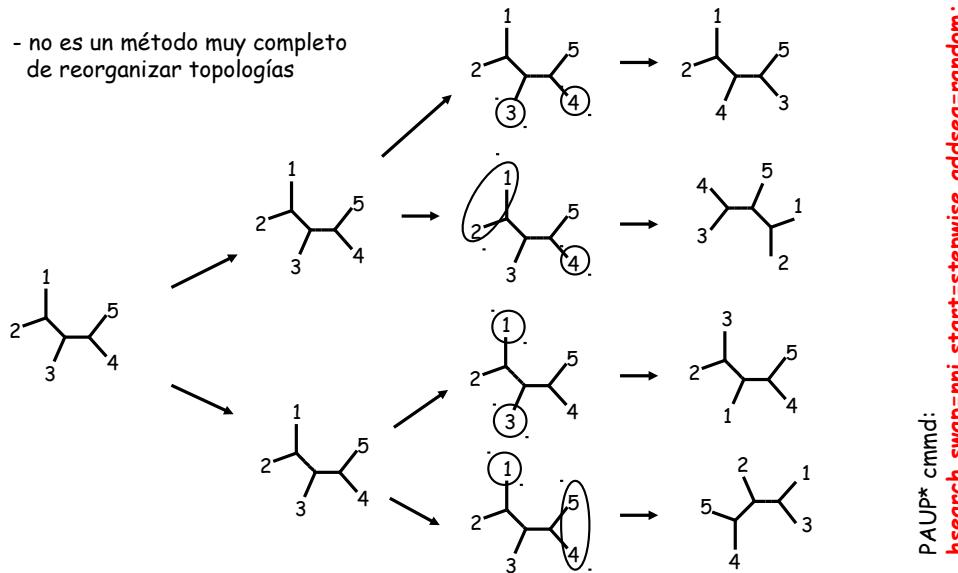
- ejemplos de perfiles de verosimilitud correspondientes a dos búsquedas con 101 y 1001 árboles semillas (1 NJ y los demás de adición secuencial aleatorizada).



Métodos heurísticos de búsqueda de árboles - intercambio de ramas (branch swapping)

- Intercambio entre vecinos más próximos (Nearest Neighbor Interchange, NNI)

- no es un método muy completo de reorganizar topologías



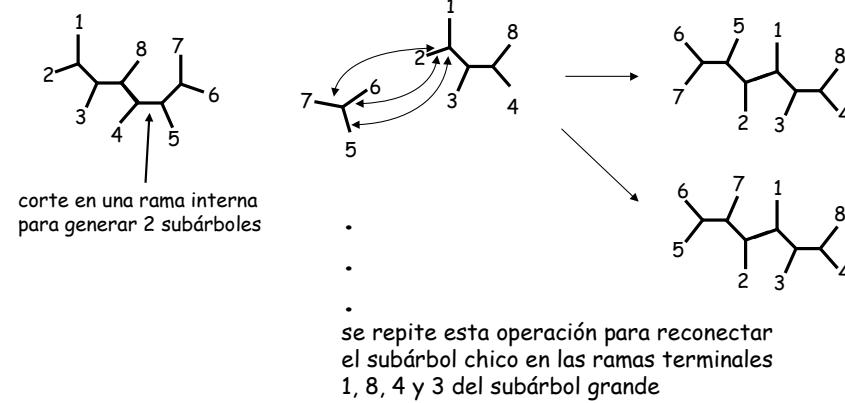
PAUP* command:
`hsearch swap=nni start=stepwise addseq=random;`

Métodos heurísticos de búsqueda de árboles - intercambio de ramas (branch swapping)

- Bisección-reconexión de árboles (Tree Bisection-Reconnection, TBR)

- Este método evalúa muchas más topols. que el NNI

se reconectan los dos subárboles en todas las posiciones posibles (ej: $3 \times 5 = 15$ subarreglos en nuestro ejemplo)



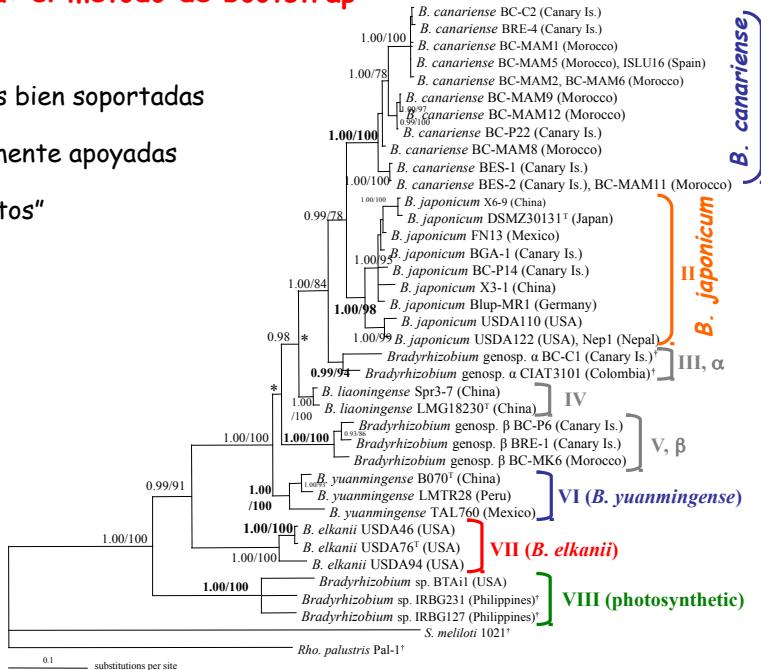
PAUP* command:
`hsearch swap=tbr start=stepwise addseq=random;`

Métodos heurísticos de búsqueda de árboles - estrategias de búsqueda para muchos OTUs $n > 25$

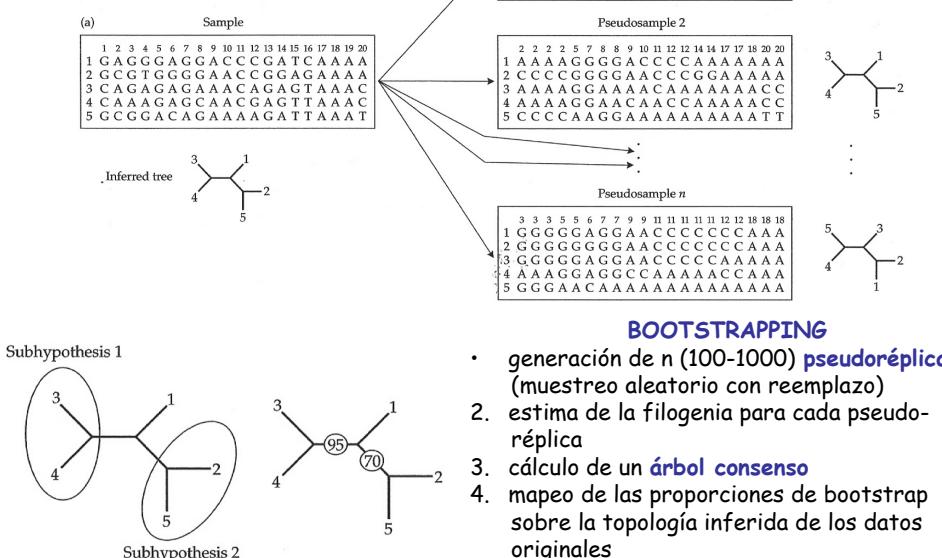
- Generalmente se combinan distintos tipos de búsquedas
 - es frecuente comenzar con (una o varias) topología generada por adición secuencial aleatorizada y mejorarla mediante un TBR
 - a veces se intercala una búsqueda NNI
- Una vez encontrada una topología mejor en una ronda de "branch-swapping", ésta sirve como topología de partida para nuevos rearreglos. Por tanto es conveniente partir de árboles "buenos" para minimizar el número de ciclos de branch swapping que se han de realizar para encontrar la topología localmente óptima. Las topologías generadas por adición secuencial aleatorizada son generalmente suficientemente "buenas" para iniciar los ciclos de branch-swapping que permiten una exploración eficiente del espacio de topologías.

Estima de la confianza que podemos tener en distintas partes de una filogenia: el método de bootstrap

"Filogenias bien soportadas
vs. pobemente apoyadas
por los datos"

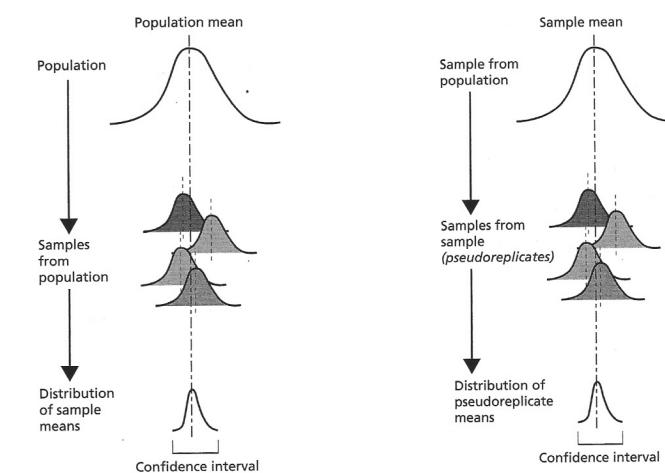


Estima del error de muestreo mediante el método de bootstrap



Estima del error de muestreo mediante el método de bootstrap

- Una vía de estimar el error de muestreo es tomar múltiples muestras de la población y comparar las estimas obtenidas de ellas. La dispersión entre estas muestras nos da una idea del error de muestreo
- El **método de bootstrap** se basa en remuestrear la propia muestra



Encuentren mucho más material en mi sitio web, ¡hasta pronto!

<http://www.ccg.unam.mx/~vinuesa/>

Welcome to Pablo Vinuesa's Research and Teaching Site

Center for Genomic Sciences (CCG - UNAM)

- environmental and evolutionary microbiology -

Main About Research Publications Group Courses - English Cursos - Español Bioinformatics Phylogeny tutorials

Contact

Pablo Vinuesa, Associate Professor
Responsible of the PhD Program in Biomedical Sciences

Centro de Ciencias Genómicas, UNAM

A.P. 565-4
Av. Universidad s/n, Col. Chamilpa,
C.P. 62210, Cuernavaca, Morelos,
MEXICO

[map](#)

Tel(s):

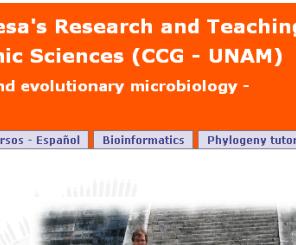
+52 777 317 5867 [\(777\) 317 5867](#), +52 777 317 5581 [\(777\) 317 5581](#)

Fax: (777) 317 5581

Red UNAM: 27691

Int. 125

[vinuesa@ccg.unam.mx](#)



NOTEWORTHY

New stuff on the site

[Taller Latinoamericano de Evolución Molecular 2011](#)

Bioinformatics

[primers4clades](#)

[primers4clades](#)

[Bioinformatics tools and resources](#)

Cursos y tutoriales

[Cursos y tutoriales en español sobre bioinformática, filogenética y evolución](#)

Apuntes más detallados sobre más temas básicos de filoinformática los encuentras aquí:

<https://github.com/vinuesa/intro2phyloinfo>

Libros de referencia recomendados:

- Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, INC., Sunderland, MA.
- Futuyma, D.J. 2005. Evolution. Sinauer Associates, INC., Sunderland, MA.
- Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland.
- Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc., NY.
- Page, R.D.M., Holmes, E.C., 1998. Molecular Evolution - A Phylogenetic Approach.
Blackwell Science Ltd, Oxford.
- Swofford, D.L., Olsen, G.J., Waddel, P.J., Hillis, D.M., 1996. Phylogenetic inference.
In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics. Sinauer Associates,
Sunderland, MA, pp. 407-514. (Una revisión excelente del campo antes de aparecer los métodos
Bayesianos)