Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

Intoducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2022, 1-5 Agosto, 2022 CCG-UNAM, Cuernavaca, Mor. México
https://github.com/vinuesa/TIB-filoinfo

Introducción a la Filoinformática: Pan-genómica y Filogenómica microbiana– NNB & CCG-UNAM, 1-5 Agosto 2022

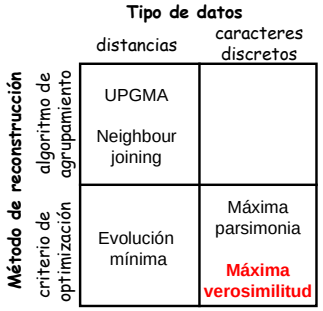
Pablo Vinuesa (vinuesa[at]ccg . unam . mx; @pvinmex)
Centro de Ciencias Genómicas, (CCG-UNAM), Campus Morelos,
Cuernavaca, México **http://www.ccg.unam.mx/~vinuesa/**

Todo el material del curso (presentaciones, tutorales y datos) lo encontrarás en:
https://github.com/vinuesa/TIB-filoinfo

• Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos de sustitución

1. El criterio de optimización de máxima verosimilitud en filogenética
1. ML y estima de parámetros del modelo de sustitución
2. ML y acomodo de la heterogeneidad de tasas de sustitución entre sitios (pInv y Gamma)
3. ML y contraste de hipótesis evolutivas (selección de modelos (LRT, AIC; hipótesis de reloj molec.)

Inferencia Filogenética y Evolución Molecular – 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)

Métodos de reconstrucción filogenética – Máxima Verosimilitud

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

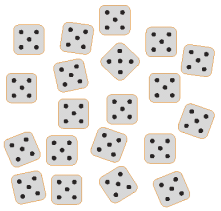
El **método de máxima verosimilitud** (**ML**) considera cada sitio variable del alineamiento (incluídos singletones). Bajo el criterio de ML se busca la topología que hace más verosímil el patrón de sustituciones de un alineamiento dado un modelo evolutivo explícito!

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



Máxima 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**, en base a la cual se evaluan las topologías alternativas. Se trata 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 condicionada a una topología particular (τ), set de longitudes de rama (υ) y modelo de sustitución (ϕ). **Es por tanto una probabilidad condicional.**
- 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.
- Hablar de la "verosimilitud de un conjunto de datos" no es correcto ya que **la verosimilitud está en 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: $L(H|D) = Pr(D|H) = Pr(D|\tau\upsilon\phi)$

Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

Intoducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2022, 1-5 Agosto, 2022 CCG-UNAM, Cuernavaca, Mor. México
https://github.com/vinuesa/TIB-filoinfo

Máxima verosimilitud y estima de parámetros de modelos de sustitución

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

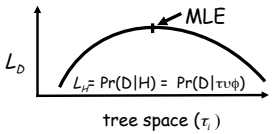
- Lo mejor es pensar en los **árboles como modelos**. La verosimilitud de una topología particular (τ) será la probabilidad de los datos dada esa topología. Cada topología tiene como parámetros las longitudes de rama (υ), y la verosimilitud de un modelo de sustitución (ϕ) cambia según varíen los valores 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 encontar 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

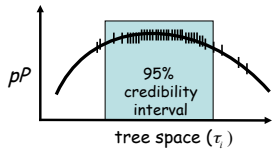
- La topología que hace de nuestros datos el resultado evolutivo más probable (dado un modelo de sust.) es la estima de máxima verosimilitud de nuestra filogenia. Por tanto, al contrario que bajo los criterios de optimización de MP, LS o ME, **bajo ML se trata de seleccionar modelos y parámetros que maximicen la función de optimización**.

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 solo 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 propabilidad posterior



Máxima 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 *rpoB* de *Bradyrhizobium japonicum* USDA110

ATGGCGCAGCAGACATTCACCGGTC

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

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

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

$\pi_A = 0.24$
$\pi_C = 0.32$
$\pi_G = 0.28$
$\pi_T = 0.16$

- 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

Máxima 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

ATGGCGCAGCAGACATTCACCGGTC

- Cálculo de ln*L* bajo el modelo de JC69

$$\ln L = 6 \ln(\pi_A) + 8 \ln(\pi_C) + 7 \ln(\pi_G) + 4 \ln(\pi_T)$$
$$= 6 \ln(0.25) + 8 \ln(0.25) + 7 \ln(0.25) + 4 \ln(0.25) = -29.1$$

- Cálculo de ln*L* bajo el modelo de F81

$$\ln L = 6 \ln(\pi_A) + 8 \ln(\pi_C) + 7 \ln(\pi_G) + 4 \ln(\pi_T)$$
$$= 6 \ln(0.24) + 8 \ln(0.32) + 7 \ln(0.28) + 4 \ln(0.16) = -26.6$$

$\pi_A = 0.24$
$\pi_C = 0.32$
$\pi_G = 0.28$
$\pi_T = 0.16$

-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.

Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

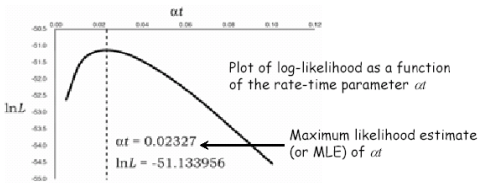
Máxima verosimilitud y estima de parámetros de modelos de sustitución

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

```
gorilla    GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan  GGACTCCTTGAGAAATAAACTGCACACACTGG
```

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

- ¿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 cuál maximiza la función.



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

Máxima verosimilitud y estima de parámetros de modelos de sustitución

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

```
*  *  *                    *  *  *
GAATCCGA ——————————————— GGATGCGT
```

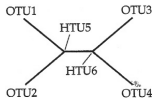
```
GAATCCGA
GGATGCGT
```

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

- 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 caracter para cada nodo interior y para cada topología !!!**.

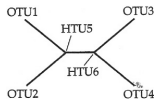
	1	2	3	4	5	6	7	8	9	... n
OTU1	A	A	G	A	C	T	T	C	A	...N
OTU2	A	G	C	C	C	T	T	C	T	...N
OTU3	A	G	A	T	A	T	C	C	A	...N
OTU4	A	G	A	G	G	T	C	C	T	...N

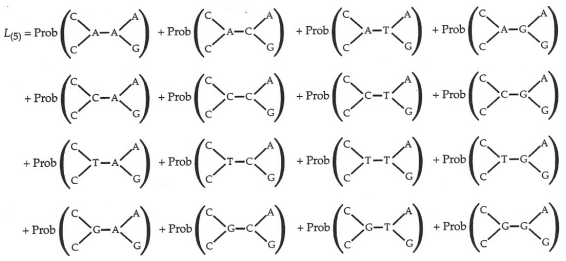


Máxima verosimilitud y estima de parámetros de modelos de sustitución

- 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





- Para 4 OTUs existen 3 topologías posibles.

Por ello hemos de repetir este cálculo para cada una de ellas con el fin de encontrar la topol. más verosímil

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

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

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

Máxima verosimilitud y estima de parámetros de modelos de sustitución

- La inferencia filogenética bajo el criterio de máxima verosimilitud implica
  MUCHISIMO TRABAJO COMPUTACIONAL
  (=> mucho tiempo de trabajo de procesador)

- Las verosimilitudes globales han de ser maximizadas para cada topol. Para ello necesitamos:
  - encontrar EMV para cada long. de rama y cada parámetro del modelo de sust.
  - ello implica calcular la verosimilitud global muchas, pero que muchas veces

- En la práctica los **árboles de ML se estiman en múltiples ciclos**, en los que se van **optimizando secuencialmente los diversos parámetros** del modelo de sustitución y longitudes de rama. La estima conjunta de todos los parámetros se hace computacionalmente prohibitiva

- Por lo general **se comienzan estos ciclos partiendo de una topología** obtenida por un método rápido, tal como **NJ o MP**. Sobre esta topología se ajustan los valores de los parámetros del modelo. A continuación se emplea algún método de reajuste de topología (branch swapping) y se ajustan las longitudes de rama, cerrando un ciclo. En múltiples ciclos consecutivos se va optimizando la topología y long. de rama, **hasta que convergen en la estima de máxima verosimilitud global**

Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

Intoducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2022, 1-5 Agosto, 2022 CCG-UNAM, Cuernavaca, Mor. México
https://github.com/vinuesa/TIB-filoinfo

Máxima verosimilitud y estima de parámetros de modelos de sustitución

¿Vale la pena tanto cómputo?

- **Uso eficiente de la información**
  - la MP ignora los sitios constantes y autapomórficos
  - los métodos de distancia pierden toda la información no capturada en la matriz de distancias pareadas
  - la inferencia bajo **MV es más consistente que** los métodos anteriores cuando existe heterogeneidad en longitudes de rama (inconsistencia en **MP**) y cuando el diámetro de los árboles es grande (inconsistencia en **métodos de distancia**)

- **Generalidad de modelo**
  - algunos modelos pueden implementarse en métodos de distancia, pero la estima del valor de los parámetros no puede hacerse de manera precisa y consistente
  - **los modelos más complejos sólo pueden implementarse bajo MV**

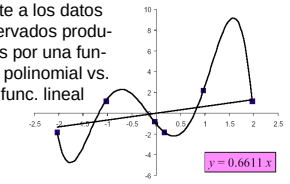
Máxima verosimilitud y estima de parámetros de modelos de sustitución

2. Selección de modelos de sustitución de secuencias de DNA

- En términos generales modelos complejos se ajustan mejor a los datos que los simples. Idealmente **se ha de seleccionar un modelo los suficientemente complejo** (rico en parámetros) como **para describir adecuadamente las características más notables del patrón de sust.** del set de datos, pero no sobreparametrizando, para **evitar colinearidad 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

Máxima verosimilitud y estima de parámetros de modelos de sustitución

2. Selección de modelos de sustitución de secuencias de DNA y Proteína

- Se deben de usar **pruebas estadísticas para seleccionar el modelo que mejor se ajusta a los datos de entre los disponibles**. Este ajuste de los modelos a los datos puede ser evaluado usando pruebas de razones de verosimilitud (likelihood ratio tests, **LRTs**) o usando criterios de información de Akaike o bayesianao (**AIC** y **BIC**, respectivamente). Se puede usar una prueba de LRT para evaluar la capacidad que tiene un modelo particular en ajustar los datos.

- Idealmente debemos de seleccionar el mejor modelo para cada gen o región genómica que queramos analizar. No conviene hacerlo para una supermatriz de alineamientos concatenados. El uso de **modelos particionados** en los que se ajusta el modelo para cada posición de los codones, por cada gen a analizar, resultan generalmente en ajustes globales significativamente mejores que **modelos promediados** para cada gen.

Máxima 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 χ^2 con q grados de libertad, donde q = diferencia entre el no. de parámetros libres entre L_1 y L_0 .**

Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

Máxima 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	-lnL
JC69	3585.54820
F81	3508.04085
HKY85	3233.34395
TrN93	3232.29439

- ¿ Qué podemos concluír 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 ?

Máxima verosimilitud y estima de parámetros de modelos de sustitución

3. Prueba de razón de verosimilitudes (LRT)

Modelo	-lnL
JC69	3585.54820
F81	3508.04085
HKY85	3233.34395
TrN93	3232.29439

H_0 a rechazar (o hipótesis anidadas a evaluar)

1. igual frec. de bases
2. Ti = Tv
3. tasas de Ti iguales

...

modelos	diff. GL = *q*	X^2	*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	← Por lo tanto el modelo seleccionado es el **HKY**
F81-TrN	5 - 3 = 2	551.4	0	
KHY-TrN	5 - 4 = 1	2.1	0.15	←

En R caluclamos el valor crítico usando nivel de certidumbre (0.95 o 0.99) y y df) :
`qchisq(0.95, df);` **rechazamos H0 si X2 observada > valor crítico**
`La p la calculamos así: 1 – pchisq(X2,df)`

http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

Máxima verosimilitud y estima de parámetros de modelos de sustitución

3. Prueba de razón de verosimilitudes (LRT)

Modelo	-lnL
HKY85	3233.34395
HKY85 +*G*	3145.29031
HKY85 +I+*G*	3142.36439

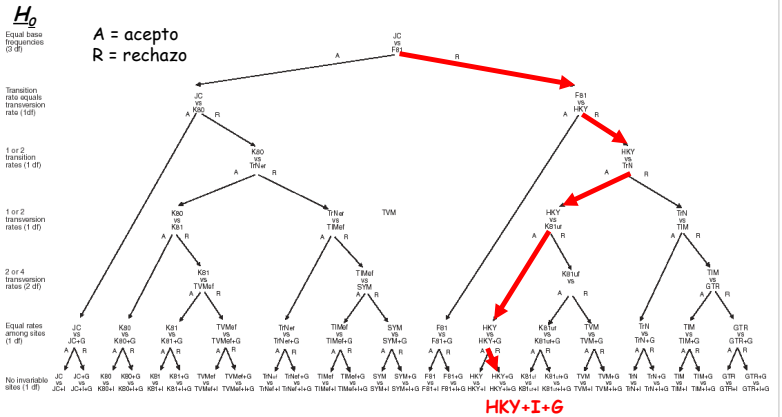
H_0 a rechazar (o hipótesis anidadas a evaluar)

1. tasa homogénea de sust entre sitios
2. no existe proporción de sitios invariantes

modelos	diff. GL = *q*	X^2	*P*	
HKY85-vs. +*G*	1	176	0	← Por lo tanto el modelo seleccionado es el **HKY+*G*** si tomamos 0.01 como punto de corte, o **HKY+I+*G*** si usamos alfa = 0.05.
HKY85+*G* vs. I+*G*	1	5.85	0.015	←

Máxima verosimilitud y estima de parámetros de modelos de sustitución

3. Esquema jerárquico de efectuar LRTs partiendo desde el modelo más sencillo (JC69)



Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

Máxima verosimilitud y estima de parámetros de modelos de sustitución

4. 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**.

Máxima verosimilitud y estima de parámetros de modelos de sustitución

4. 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**.

$AIC_i = -2\ln L_i + 2 N_i$ N_i = no. de parámetros libres en el modelo *i*

L_i = verosimilitud bajo el modelo *i*

Máxima verosimilitud y estima de parámetros de modelos de sustitución

4. 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 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 modelos cadidatos. 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 **Δ_i > 10 carecen de soporte**.

Máxima verosimilitud y estima de parámetros de modelos de sustitución

4. 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:

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

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.

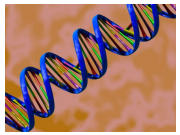
Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

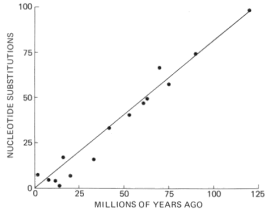
Máxima verosimilitud y estima de parámetros de modelos de sustitución

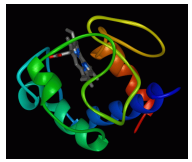
5. LRT del reloj molecular global

Emile Zuckerkandl, uno de los pioneros de la disciplina de la **evolución molecular**, fue de los primeros en descubrir que las moléculas de **DNA y las proteínas** que codifican son "**documentos de la historia evolutiva**" dada la relativa constancia con la que acumulan variaciones (mutaciones).

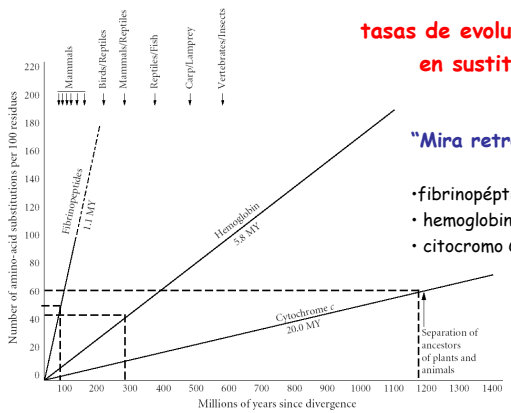
- Nace la **hipótesis del reloj molecular** que postula que la distancia genética entre dos secuencias que codifican para la misma proteína, pero aislada de diferentes organismos, incrementa más o menos linealmente con respecto al tiempo de divergencia entre las especies. Es decir, la tasa de evolución para una proteína dada es relativamente contante a lo largo del tiempo







tasas de evolución de tres proteínas en sustituciones/sitio/MY



"Mira retrospectiva en el tiempo"

- fibrinopéptidos: 50- 200 MY
- hemoglobinas: 200- 800 MY
- citocromo *C*: 400-1300 MY

- Distintas proteínas presenta diversas tasas de sustitución. Así los fibrinopéptidos presentan relativamente pocas constricciones, presentando una elevada tasa de sustitución neutral. Citocromo *C*, en cambio, presenta mayores constricciones evolutivas y presenta una tasa de sustitución menor. La hipótesis del reloj molecular dice que esta tasa, para ciertas proteínas, se mantiene constante en distintos linajes.
(de Hartl y Clark, 1997. Principles of Population Genetics, Sinauer)

Máxima verosimilitud y estima de parámetros de modelos de sustitución

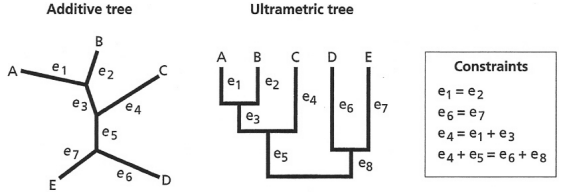
5. LRT del reloj molecular global

- En ausencia de un reloj molecular (**modelo de tasas libres de evolución**) y para un árbol bifurcante no enraizado se han de inferir **2n-3 longitudes de rama**.

- Forzando un modelo de evolución **bajo reloj molecular global** (todos los linajes evolucionan con igual tasa), el árbol está enraizado y se tienen que estimar sólo **n-1 longitudes de rama**.

(2n-3) = 7 ramas independientes

(n-1) = 4 ramas independientes



Máxima verosimilitud y estima de parámetros de modelos de sustitución

5. LRT del reloj molecular global y local

- La hipótesis de **reloj molecular** = H_0 por representar un caso particular (restringido) de la hipótesis más general (H_1) que asume distintas tasas para cada linaje. Por ello podemos evaluar la hipótesis de reloj molecular usando LRTs. Para ello comparamos el valor del estadístico **LRT** con una distribución X^2 con (2n-3) – (n-1) = **n-2 GL**, ya que la única diferencia en estima de parámetros es el número de longitudes de rama que hay que estimar.

- La **hipótesis de reloj molecular puede ser relajada**, permitiendo una tasa constante de sustitución sólo dentro de un clado particular y asumiendo tasas heterogéneas para el resto de los linajes de una filogenia (modelo de **reloj molecular local**).
El reloj molecular global sería un caso especial de un modelo de reloj local, el cual a su vez es representa un modelo constreñido o especial del modelo de tasas libres. Estos modelos pueden ser comparados de manera pareada usando LRTs.

Tema 6: Máxima verosimilitud: estima de parámetros y selección de modelos

Máxima verosimilitud y estima de parámetros de modelos de sustitución

5. LRT del reloj molecular global: un ejemplo

1. Usaremos nuevamente el set de datos de mtDNA-primates.nex, eliminando a Lemur_catta, Tarsius_syrichta y Saimiri_scireus y excluyendo regiones no codificadoras.
2. Obtenemos una filogenia NJ sobre la cual estimaremos los parámetros del modelo de sust. HKY+I+G bajo modelo de tasas libres.
3. Enraizamos el árbol en memoria con **rootTrees** y reestimamos el valor global de verosimilitud usando los mismos valores de parámetros estimados en (2), forzando el modelo de reloj molecular global (**lset clock=yes**;)

-lnL	hipótesis
3124.46376	tasas libres
3126.86992	reloj global

$\Delta = 2(\ln L_1 - \ln L_0) = 4.81232$

GL = no. OTUs -2 = 9-2 = 7

P = 0.6828

podemos aceptar la hipótesis de reloj global