Clase 3

Soporte y modelos estadísticos

Bootstrapping

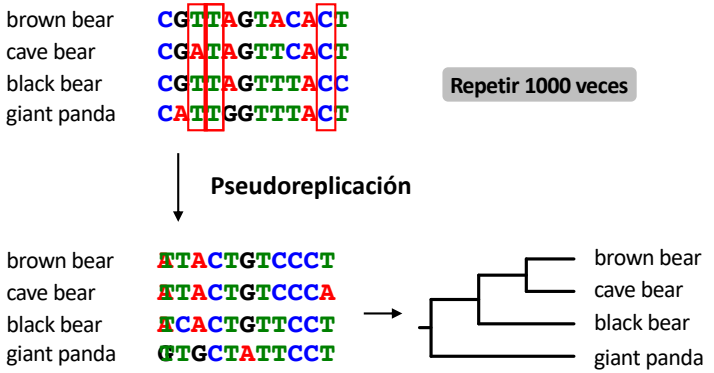
El bootstrap no paramétrico

- La incertidumbre en el estimativo del árbol se pude inferir indirectamente usando **análisis de bootstrap**
- "uno se sube a sí mismo usando sus propios bootstraps"

  - El análisis usando Bootstrap se puede usar en varios métodos filogenéticos:
    - Máxima parsimonia
    - Métodos basados en matrices de distancia
    - Máxima verosimilitud

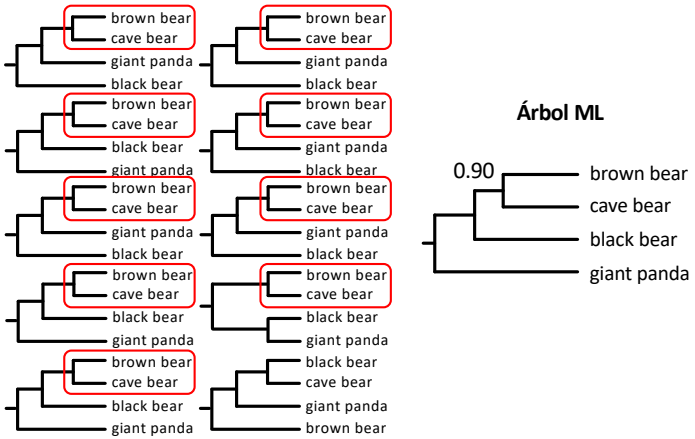
3

Bootstrap



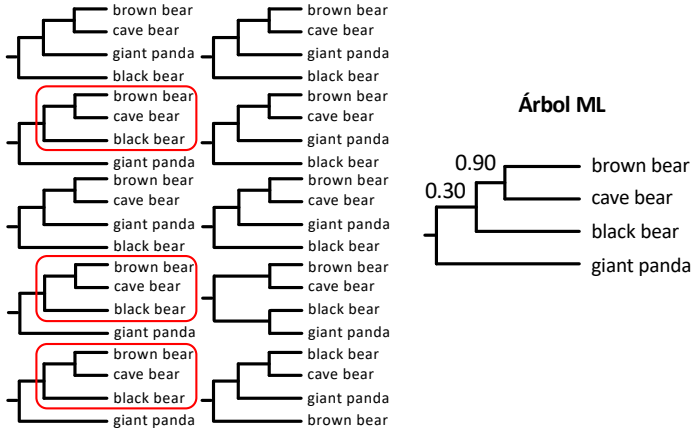
4

Bootstrap



5

Bootstrap



6

Interpretar valores de bootstrap

- **Felsenstein (1985)**
  El bootstrap nos da un intervalo de confianza que contiene *la filogenia que sería estimada por muestrear repetidamente caracteres de la distribución existente*

- Los valores del Bootstrap son **medidas de repetibilidad**
  - Es alto cuando hay muchos datos disponibles
  - Tiene poco significado cuando datos del genoma completo están disponibles

Soltis & Soltis (2003) *Stat Sci*

7

Métodos filogenéticos comunes

1. Máxima parsimonia
2. Métodos de distancia
3. Máxima verosimilitud
4. Inferencia Bayesiana

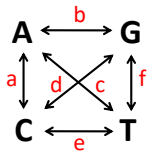
Métodos estadísticos

8

Modelos de sustitución

Modelos de sustitución de nucleótidos

Matriz de tasas



Frecuencias de las bases

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

JC	HKY	GTR
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$

10

Variación de tasas en sitios

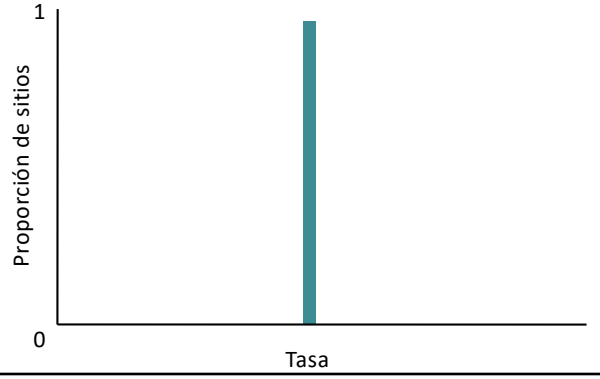
```
CTAT--GGCACCCAGCCCATGCAT-GGT
CTAA--GGCAACCAGCCCATACAT-GCT
CTATGTGGCAACCAGCCCATGCAT-GCT
ATATGTGGCAGCCAG-----GCATAGGT
ATATGTGGCAGCCAGCCCATGCATAGGT
```

Medio Lento Rápido

11

Variación de tasas en sitios

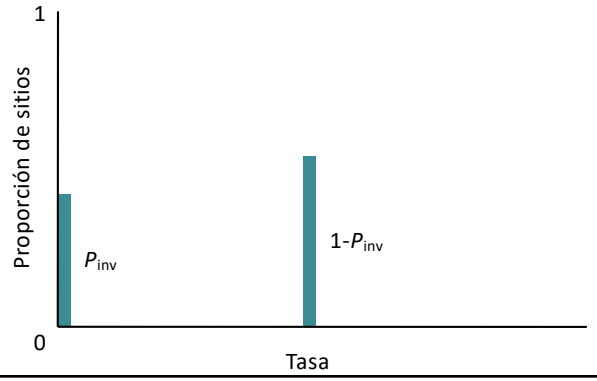
- Igual entre sitios



12

Variación de tasas en sitios

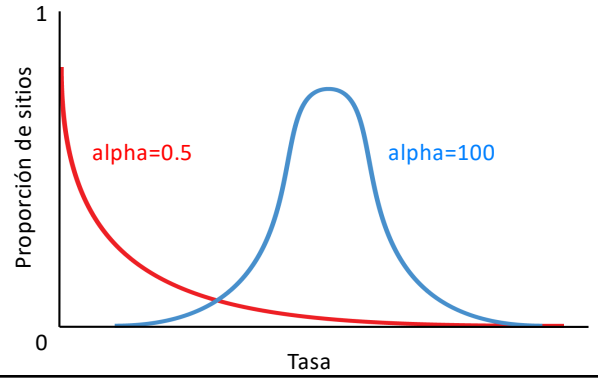
- Proporción de sitios invariables (modelos **+I**)



13

Variación de tasas en sitios

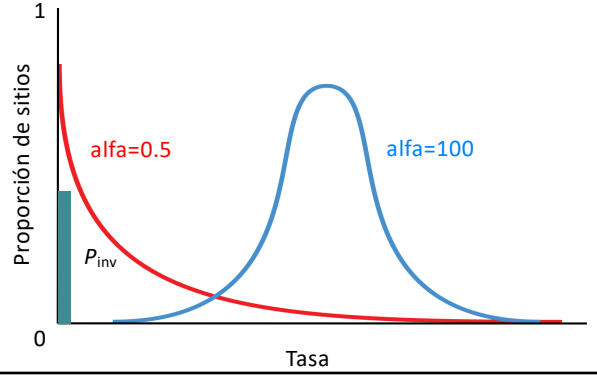
- Variación de tasas entre sitios con distribución gama (modelos **+G**)



14

Variación de tasas en sitios

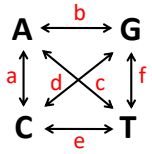
- Tasas entre sitios con distribución gama y una proporción de sitios invariables (modelos **+G+I**)



15

Modelos de sustitución de nucleótidos

Matriz de tasas



Frecuencias de las bases

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

Tasas de sitios

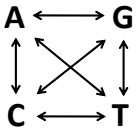
+ I + G

JC	**HKY**	**GTR**	**GTR+I+G**
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	π_A, π_C, π_G, π_T	π_A, π_C, π_G, π_T	π_A, π_C, π_G, π_T
			I, G

16

Modelos de sustitución de nucleótidos

Matriz de tasas



Frecuencias de bases

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

Tasas de sitios

+ I + G

Número de modelos

203 x 15 x 4 = 12,180

En filogenética solo exploramos un pequeño grupo de estos

17

Proporción de sitios invariables

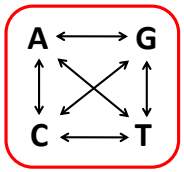
- A menudo se sobre-estiman cuando se analisan especies
- No distinguen:
  - Sitios que son **invariables** y no pueden cambiar
  - Sitios que son **constantes** y por razones estocasticas no han cambiado
- Tiene poco significado biológico
- Los sitios lentos se pueden describer bien usando **+G**

Usamos modelos +G models para tomar en cuenta la variación en la tasa entre sitios

18

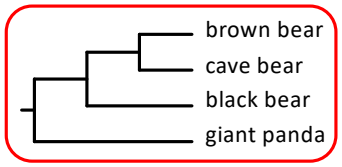
Asunciones fundamentales

Reversible



Estacionario

$\pi_A + \pi_C + \pi_G + \pi_T = 1$



CGTTAGTACACT
CGATAGTTCACT
CGTTAGTTTACC
CATTGGTTTACT

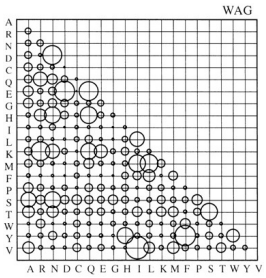
Homogéneo

Sitios independientes

19

Matrices de sustitución de amino ácidos

- Matriz de probabilidades de sustitución de 20x20
- Demasiados parámetros para estimar
  - Modelo GTR para ADN: 6 parámetros
  - Modelo GTR para proteínas: 190 parámetros
- Se estiman probabilidades de sustitución usando cantidades grandes de datos
  - PAM
  - BLOSUM
  - JTT
  - WAG



20

Selección de modelos

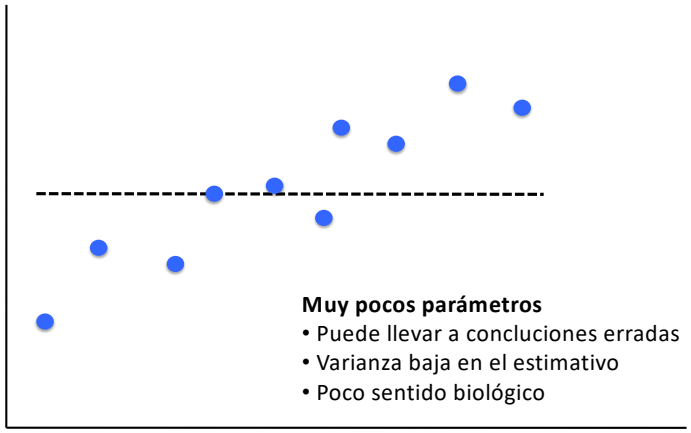
Selección de modelos

1. **Selección subjetiva de modelos**
   - Eligiendo un modelo que parezca sensato
   - Balanceando el número de parámetros contra la cantidad de datos disponible
   - Motivación biológica

2. **Selección objetiva de modelos**
   - Usando teoría de la infromación y hacerlo a computador
   - Motivación estadística

22

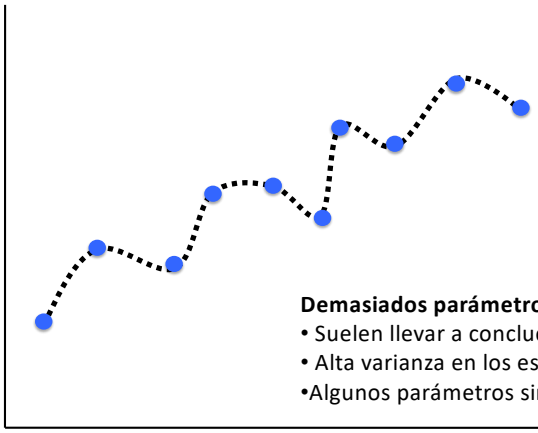
Selección de modelos



Muy pocos parámetros
- Puede llevar a concluciones erradas
- Varianza baja en el estimativo
- Poco sentido biológico

23

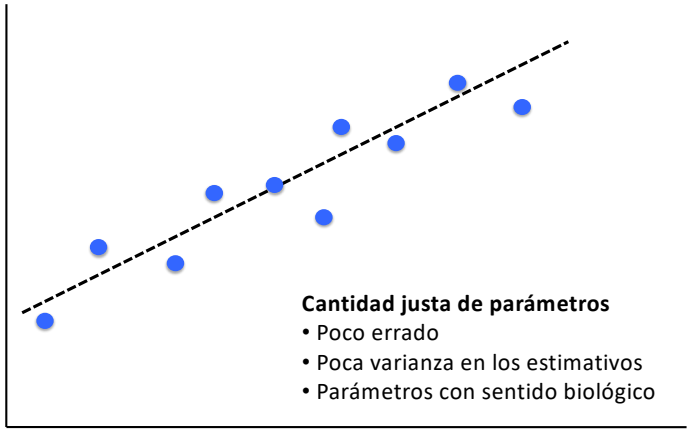
Selección de modelos



Demasiados parámetros
- Suelen llevar a concluciones acertadas
- Alta varianza en los estimativos
- Algunos parámetros sin sentido biológico

24

Selección de modelos



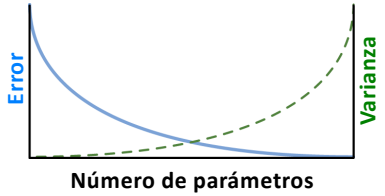
Cantidad justa de parámetros
- Poco errado
- Poca varianza en los estimativos
- Parámetros con sentido biológico

25

Selección de modelos

- Adicionar parámetros *siempre* mejora el ajuste del modelo a los datos
- Pero adicionar parametros lleva a más varianza en sus estimativos

¿Vale la pena mejorar el ajuste dado el costo de más parametros?



26

Selección de modelos

- **Test de la proporción de verosimilitud, *Likelihood-ratio test* (LRT)**
  Usado para comparar modelos anidados
- **Criterio de información de Akaike (AIC)**
  AIC = -2ln(likelihood) + 2*k*
- **Criterio de información Bayesiano (BIC)**
  BIC = -2ln(likelihood) + *k*ln(*n*)

27

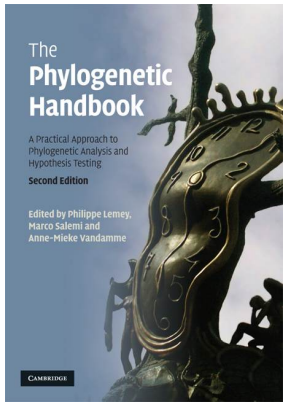
Modelos de sustitución en la práctica

- El árbol filogenético es un parámetro áltamente robusto al modelo usado
- **GTR+G** es acceptable para la mayoría de los datos

28

Referencias útiles

- **Model selection in phylogenetics**
  Sullivan & Joyce (2005) *Annual Review of Ecology, Evolution, and Systematics*, 36: 445–466.
- **The effects of partitioning on phylogenetic inference**
  Kainer & Lanfear (2015) *Molecular Biology and Evolution*, 32: 1611–1627.



29