Clase 2

Datos y métodos filogeneticos

Datos filogenéticos

1. **Preparación de datos**
   - Muestreo de grupos taxonómicos y regiones genómicas
   - Alineamiento
   - Filtración de datos
2. **Inferencia filogenética**
   - Selección de modelos
   - Estimación de parametros (incluyendo el árbol)
   - Análisis e interpretación adicional

2

Datos filogenéticos

- **Seleccionar datos para optimisar señal:ruido**
  - Regiones que evolucionan lento para sistemas antiguas
  - Regiones que evolucionan rápido para sistemas recientes
- **Homoplasia**
  - Organismos comparten similitudes que no reflejan su evolución
- **Aprovechar los recursos disponibles**

NCBI DRYAD TreeBASE

3

Típos de datos

- **Secuencias**
  - Nucleótidos
  - Amino ácidos
- **Datos binarios** (presencia/ausencia de características genomicas)
- **Microsatelites** (número de repeticiones)
- **Polimorfismos de un solo nucleotido** (*Single Nucleotide Polimorphisms,* SNPs)
- **Sequencias de representación reducida**

4

Secuencias

- **Codificantes**
  - ARN ribosómico
  - Codificantes de proteínas
- **No-codificantes**
  - Regiones intergenicas
  - Intrones
- **Amino ácidos**

5

Secuencias

Gen codificante

```
                  M  R  E  P  Y  S  R
brown bear    CGTTAG--CATGAGGGAACCCTACTCTAGG
                  M  R  E  P  Y  S  R
cave bear     CGATAG-TCATGAGGGAACCCTACTCTAGG
                  M  R  E  S  Y  P  R
black bear    CGTTAG-TTATGAGGGAATCCTACCCTAGG
                  M  R  H  S  -  S  R
panda         CA--GGTTTATGAGGCATTCC---TCTAGG
```

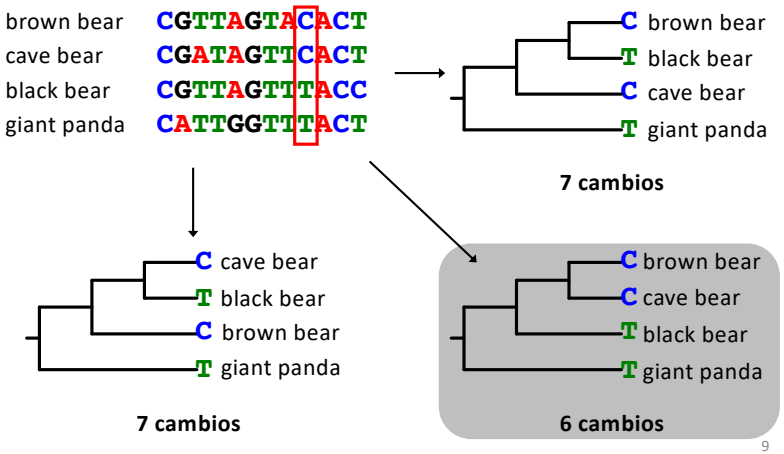
6

Clase 2.2

Métodos filogenéticos

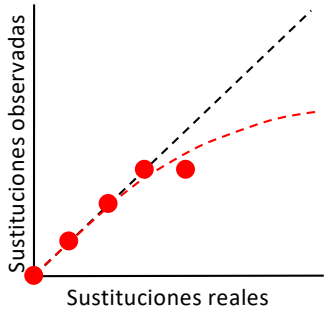
Máxima parsimonia

Máxima parsimonia



Máxima parsimonia

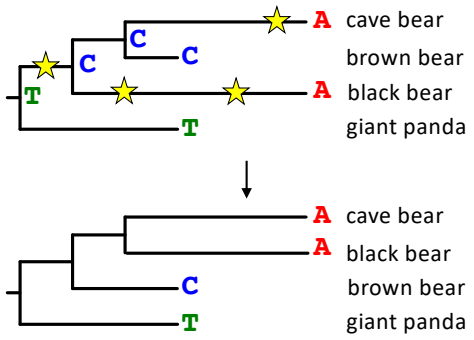
- Identifica la topología que expliquen los datos genéticos con el mínimo numero possible de cambios evolutivos
- Frecuentemente usada para el análisis de datos morfológicos
- Hoy es usada rara vez para el analysis de datos moleculares
  - No permite estimar tasas moleculares o tiempos de divergencia
  - Tiene efectos indeseados cuando hay multiples sustituciones moleculares



A A A A A
A T T T T
C C G G G
A A A A A
T T T T T
T T T T T
A A A A A
G G G G G
T T T A C

- La máxima parsimonia no toma en cuenta multiples sustituciones en el mismo sitio del alineamiento
- Esto lleva a un problema llamado la **atracción de ramas largas**
  - Ramas largas = multiples sustituciones
  - Similitudes (homoplasia) emerge estocásticamente
  - Las ramas largas se agrupan

Atracción de ramas largas



Podemos corregir multples eventos usando modelos estadísticos

Métodos filogenéticos populares

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

Métodos estadísticos

13

Máxima verosimilitud

Verosimilitud filogenética

Probabilidad	Modelo
Árbol 1	0.1
Árbol 2	0.7
Árbol 3	0.15
Árbol 4	0.05
Suma	1

15

Verosimilitud filogenética

Probabilidad	Modelo
Árbol 1	0.1
Árbol 2	0.7
Árbol 3	0.15
Árbol 4	0.05
Suma	1

Una función matemática nos da la probabilidad de cáda árbol:

La función de verosimilitud filogenética

16

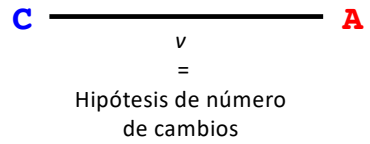
Verosimilitud filogenética

- Una sustitución molecular es un evento estocástico

17

Verosimilitud filogenética

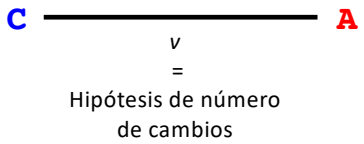
- Una sustitución molecular es un evento estocástico
  - **Nos interesa la probabilidad de transición**



18

Verosimilitud filogenética

- Una sustitución molecular es un evento estocástico
  - **Nos interesa la probabilidad de transición**

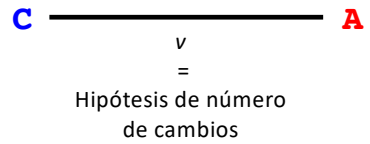


- La **distribución poisson** describe procesos estocasticos

19

Verosimilitud filogenética

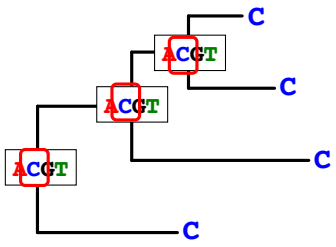
- Una sustitución molecular es un evento estocástico
  - **Nos interesa la probabilidad de transición**



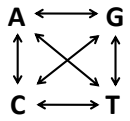
- La **distribución poisson** describe procesos estocasticos
  - La probabilidad de transición es dada por la ecuación e^{Qv}

20

La verosimilitud de una hipótesis



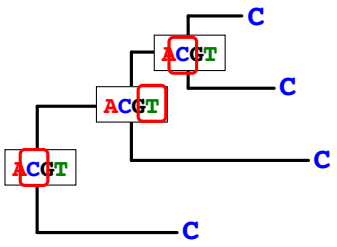
cave bear	CGTTAGTACACT
black bear	CGATAGTTCACT
brown bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT



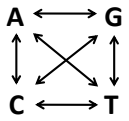
La verosimilitud se multiplica para todos los sitios

Hay una probabilidad muy baja de observar cualquier alineamiento específico

La verosimilitud de una hipótesis



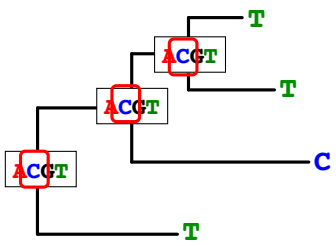
cave bear	CGTTAGTACACT
black bear	CGATAGTTCACT
brown bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT



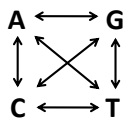
La verosimilitud se multiplica para todos los sitios

Hay una probabilidad muy baja de observar cualquier alineamiento específico

La verosimilitud de una hipótesis



cave bear	CGTTAGTACACT
black bear	CGATAGTTCACT
brown bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

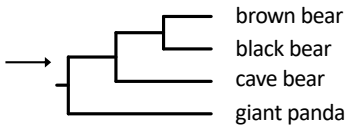


La verosimilitud se multiplica para todos los sitios

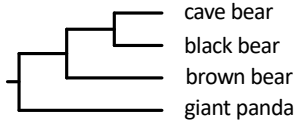
Hay una probabilidad muy baja de observar cualquier alineamiento específico

Máxima verosimilitud

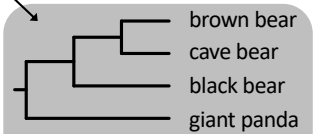
brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT



lnL = -1203.83



lnL = -1241.47



lnL = -908.58

Optimización de la verosimilitud

- Buscar en el espacio de posibles árboles y parametros
- Calcular la verosimilitud de cada uno
- Encontrar el caso con la major verosimilitud
- Optimización de multiples variables

25

Cómo encontrar el major árbol

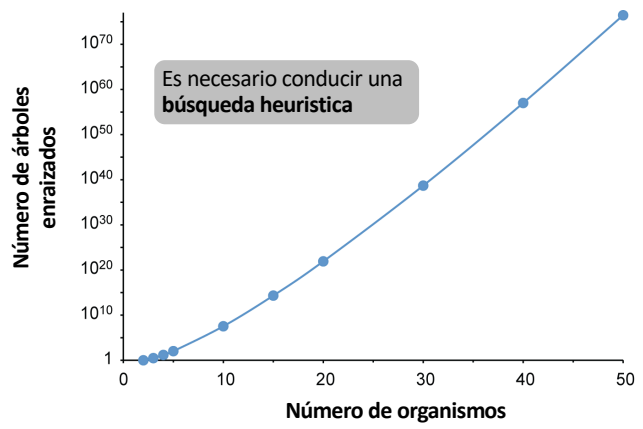
- Para *n* organismos, el número de posibles árboles no-enraizados (B_n) es:

$$B_n = 1 \times 3 \times 5 \times \ldots \times (2n-5) = \prod_{i=3}^{n} (2i-5)$$

- Por ejemplo:
  - 4 organismos → 3 árboles
  - 5 organismos → 15 árboles
  - 10 organismos → 2,027,025 árboles

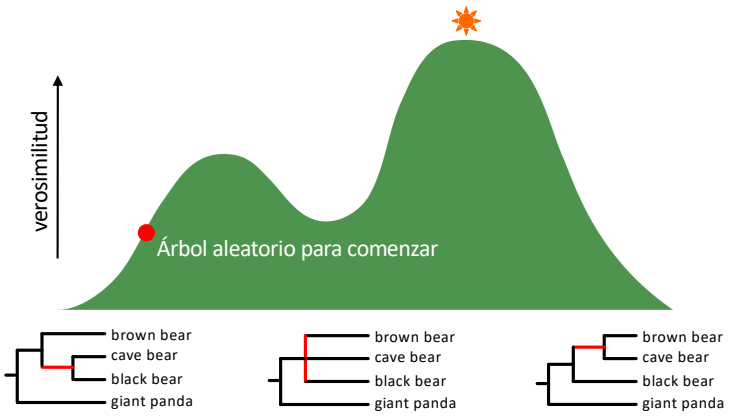
26

Cómo encontrar el major árbol



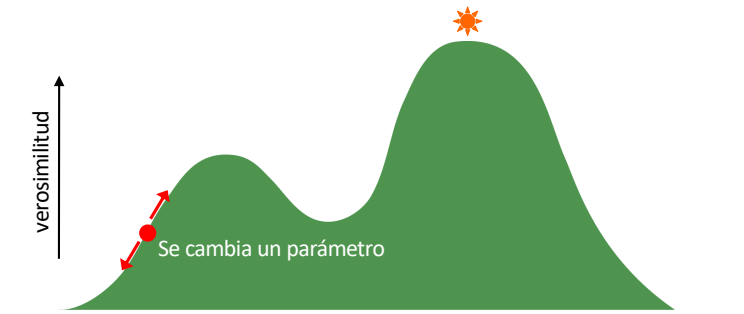
27

Búsqueda heuristica



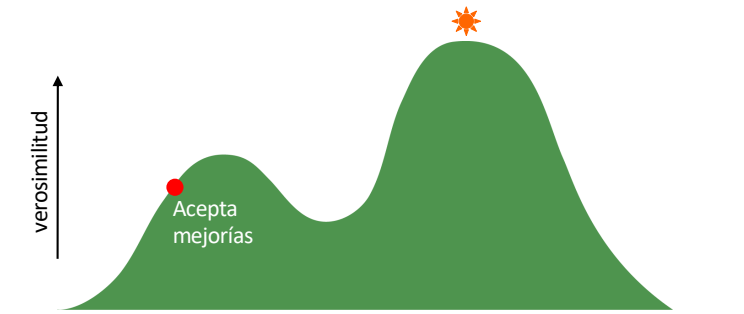
28

Búsqueda heuristica



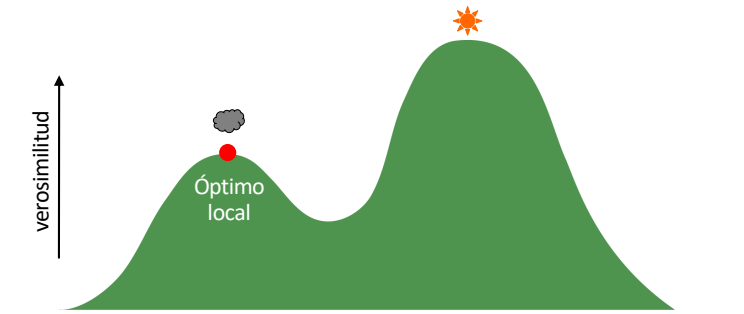
29

Búsqueda heuristica



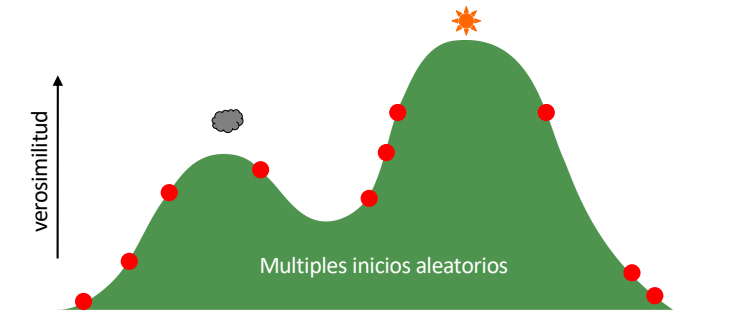
30

Búsqueda heuristica



31

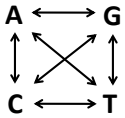
Búsqueda heuristica



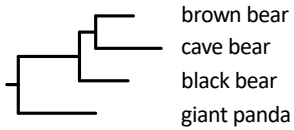
32

Estimativos en máxima verosimilitud

Parametros para un modelo de sustitución



Un árbol filogenético con longitudes de ramas



33

¿Por qué la palabra *verosimilitud*?

Verosimilitudes son caracteristicas de modelos

Probabilidad	Modelo 1
Árbol 1	0.1
Árbol 2	0.7
Árbol 3	0.15
Árbol 4	0.05
Suma	1

34

¿Por qué la palabra *verosimilitud*?

Verosimilitudes son caracteristicas de modelos

Probabilidad	Modelo 1	Modelo 2	Modelo 3
Árbol 1	0.1	0.2	0.05
Árbol 2	0.7	0.29	0.35
Árbol 3	0.15	0.5	0.4
Árbol 4	0.05	0.01	0.2
Suma	1	1	1

35

¿Por qué la palabra *verosimilitud*?

Verosimilitudes son caracteristicas de modelos
No suman a 1

Probabilidad	Modelo 1	Modelo 2	Modelo 3
Árbol 1	0.1	0.2	0.05
Árbol 2	0.7	0.29	0.35
Árbol 3	0.15	0.5	0.4
Árbol 4	0.05	0.01	0.2
Suma	1	1	1

$P(D|H)$

36

¿Por qué la palabra *verosimilitud*?

Verosimilitudes son caracteristicas de modelos
No suman a 1

Probabilidad	Modelo 1	Modelo 2	Modelo 3
Árbol 1	0.1	0.2	0.05
Árbol 2	0.7	0.29	0.35
Árbol 3	0.15	0.5	0.4
Árbol 4	0.05	0.01	0.2
Suma	1	1	1

Probabilidad es una
caracteristica de los datos
Suma a 1

37

Fortalezas y debilidades

- **Fortalezas**
  - Es un método estadistico riguroso
  - Corrige multiples sustituciones y atracción de ramas largas
  - Robusto a violaciones de sus asunciones
- **Debilidades**
  - Es difícil usar modelos con muchos parámetros
  - Puede ser difícil cubrir el espacio de posibles árboles

38

Programas

RAxML 

PhyML 

MEGA 

PAML

IQ-TREE

39

Métodos filogenéticos en la práctica

- **Máxima parsimonia**
  - Usada frequentemente para análisis de datos morfológicos
  - Rara vez usada para análisis molecular
- **Máxima verosimilitud**
  - Ampliamente utilizada, pero ahora parcialmente reemplazada por métodos Bayesianos

40