Curso fundamenteal de Inferencia Filogenética Molecular

Pablo Vinuesa (vinuesa@ccg.unam.mx)

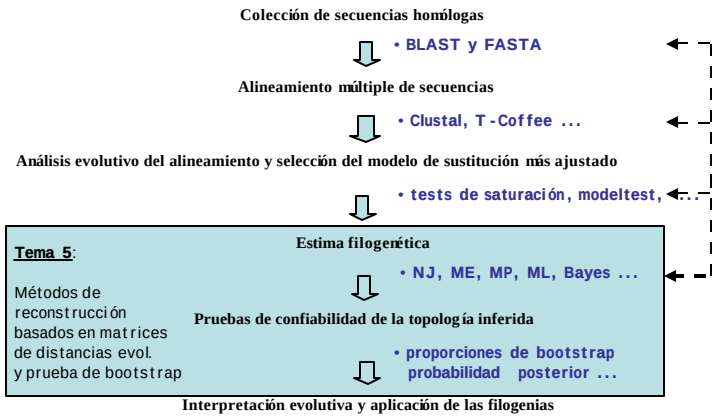
Progama de Ingeniería Genómica, CCG, UNAM

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

Tutor: PDCBM, Ciencias Biológicas, PDCBioq. y
Profesor de la Lic. Ciencias Genómicas y posgrado

- **Tema 5: Métodos de distancias y prueba de bootstrap**
  - **clasificación de métodos filogenéticos** y **tipos de datos**
  - **distancias evolutivas** y **patrísticas**
  - métodos de distancias basados en **criterios de optimización: mínimos cuadrados y evolución mínima**
  - métodos de distancias basados en **algoritmos de agrupamiento: UPGMA y NJ**
  - estima del error de muestreo: prueba de **bootstrap**

Protocolo básico para un análisis filogenético de secuencias moleculares



Inferencia filogenética molecular – clasificacón 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 método de búsqueda basado en un criterio de optimización** para encontrar la topología óptima bajo el criterio seleccionado

Tipo de datos

Método de reconstrucción	distancias	caracteres discretos
algoritmo de agrupamiento	UPGMA y Neighbor joining	×
criterio de optimización	Mínimos cuadrados y Evolución mínima	Máxima parsimonia y Máxima verosimilitud

Inferencia filogenética molecular – Métodos de distancia

- **Tipos de datos:**
  - **caracteres**: proveen información sobrecada OTU individual
  - **distancias**: cuantificación de la dis-similitud entre pares de OTUs

- **Caracter:** (característica o variable independiente bien definida que en un OTU puede presentar dos o más estados mutuamente excluyentes; **estados de caracter**)
  - **cuantitativos** (est. de car. generalmentecontínuos; ej. altura)
  - **cualitativos** (est. de car. discretos; binarios o multiestado; gralte. revesibles)

- **Evolución de caracteres:**
  Los métodos de reconstrucción filogenética requieren que se hagan suposiciones explícitas sobre:
  1.- no. de pasos discretos necesarios para que se dé un cambio en estado de caracter
  2.- la probabilidad con la que acontece un cambio en estado de caracter

- **Direccionalidad en la evolución de los cambios de estado de caracter (EC):**
  - **caracteres ordenados**: siguen secuencia específica de pasos (matrices de pasos)
  - **caracteres desordenados**: los cambios en EC se dan en un solo paso (nt)

Inferencia filogenética molecular – Métodos de distancia

- **Datos de distancia**:

  - siempre involucran la **comparación entre pares de OTUs**

  - la mayor parte de los métodos moleculares generan datos de caracteres; éstos han de ser transformados en distancias para poder ser analizados por métodos basados en matrices de distancias (p. ej. NJ, UPGMA, EM)

- **¿Porqué transformar caracteres en distancias?**

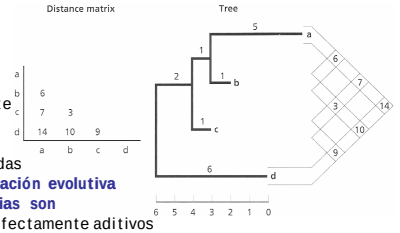
1.- Una larga lista de estados de caracter, como una secuencia de DNA ó aa, carece en sí misma de significado evolutivo; en cambio, decir que 3 secuencias A <-> B <-> C presentan 95% y 50% de identidad entre ellas evoca una imagen intuitiva del "grado de parentesco"

2.- Los modelos de sust. de secuencias corrigen posibles múltiples sustituciones; estas correcciones se aplican a las distancias pero no a las secuencias (o datos)

3.- Los métodos de reconstrucción basados en matrices de dist. son muy rápidos

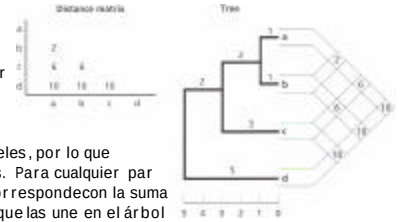
Inferencia filogenética molecular – métodos de distancias

- **Distancias topológicas**



- Las **distancias aditivas o métricas** definen a una **topología aditiva**
El árbol métrico representa perfectamente a las distancias aditivas. Nótese que las secs. b y c son las más similares [*d* (b,c) = 3], pero no son las más relacionadas evolutivamente. **El nivel de similitud y relación evolutiva coincidirán sólamente cuando las distancias son ultramétricas**. Datos reales nunca son perfectamente aditivos

- Las **distancias ultramétricas** definen una **topología ultramétrica**.
Biológicamente dist. ultram. se ajustan a un árbol enraizado bajo el reloj molecular
La sec. d es equidistante a todas las demás y la sec. c es equidist. de a y b.
Si tomamos 3 secs. cualesquiera, las dist. entre ellas definen un triángulo isósceles, por lo que las distancias mostradas son ultramétricas. Para cualquier par de secs, el valor de dist. en la matriz se corresponde con la suma de long. de ramas en el caminio más corto que las une en el árbol



Inferencia filogenética molecular – métodos basados en matrices de distancias

- En un mundo perfecto, las distancias evolutivas estimadas serían perfectamente aditivas, en cuyo caso podríamos encontrar una combinación de long. de ramas (a, b, c, d, e) tales que el camino a través del árbol conectando el OTU i con el j (p_{ij} = **distancia topológica o patrística**) reflejaría exactamente la **distancia evolutiva** correspondiente (d_{ij}).
Pero "el mundo" (homoplasias) y los métodos no son perfectos ...

  - De ahí que existan 2 estrategias que buscan minimizar el desfase entre la distancia evolutiva y la distancia topológica y por lo tanto representan criterios de optimización:

    1. **métodos de "bondad de ajuste":** buscan el árbol métrico que mejor acomoda las distancias "observadas" usando el método de **mínimos cuadrados**

    2. **métods de evolución mínima:** buscan el **árbol cuya suma de longitudes de rama es la mínima**

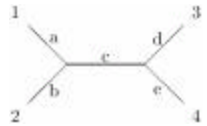
Inferencia filogenética molecular – métodos basados en matrices de distancias

- **Método de los mínimos cuadrados (medidas de la "bondad de ajuste")**

$$SS = \sum_{i<j} \frac{(d_{ij} - p_{ij})^2}{d_{ij}^k}$$

El **método de los mínimos cuadrados** permite encontrar la combinación de valores de (*a*, b, *c*, *d* y *e*) que **maximiza el ajuste entre p_{ij} y d_{ij}**. Encontrar las long. de ramas mejor ajustadas implica minimizar la suma ponderada de cuadrados.

$w = 1/ d^K_{ij}$ representa un **factor de ponderación** inversamente proporcional a la distancia estimada, donde $k = 0$ ó $k = 2$. Así las divergencias profundas tienen menor peso que las más recientes, las cuales se pueden estimar mejor.



	1	2	3	4
1		$p_{12} = a + b$	$p_{13} = a + c + d$	$p_{14} = a + c + e$
2	d_{12}		$p_{23} = b + c + d$	$p_{24} = b + c + e$
3	d_{13}	d_{23}		$p_{34} = d + e$
4	d_{14}	d_{24}	d_{34}	

diag. super.: dist. patrísticas

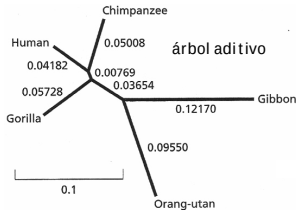
diag. infer.: dist. evolutivas

Inferencia filogenética molecular – métodos basados en matrices de distancias

• **Método de los mínimos cuadrados** (medidas de la "bondad de ajuste")

	Human	Chimp	Gorilla	Orang-utan	Gibbon
Human	–	0.09190	0.1083	0.1790	0.2057
Chimp	0.0919	–	0.1134	0.1940	0.2168
Gorilla	*0.1068*	**0.1151**	–	0.1882	0.2170
Orang-utan	**0.1816**	*0.1898*	**0.1893**	–	0.2172
Gibbon	**0.2078**	*0.2160*	*0.2155*	0.2172	–

Distancias K2P (sobre la diagonal) y distancias topológicas obtenidas por CM para mtDNAs. En **negritas dt > dg**; en *cursiva dt < de* (dt=dist. topol.; dg= dist. observada o genética)



• Las **dt > dg** pueden explicarse por homoplasias en algunas ramas

• Las *dt < dg* no pueden explicarse fácilmente y son contra-intuitivas, ya que implicarían que aconteció menos cambio evolutivo que el observado!

• Ello ha llevadoa algunos investigadores a criticar fuertemente el método de los MC para estimar la long. de las ramas

Inferencia filogenética molecular – métodos basados en matrices de distancias

• Criterio de optimización de **Evolución Mínima**

- dada una **topología aditiva** para *n* secuencias, existen **(2*n* - 3) ramas**, cada una con una longitud l_i. La suma de estas long. de ramas es la **longitud *L* del árbol**:

$$L = \sum_{i=1}^{2n-3} e_i$$

- dados dos árboles, aquel que minimiza la suma de longitudes de ramas ***L*** (estimadas por MC) es el mejor según el criterio de **EM**

• El criterio de optimización de **EM** es por tanto similar al de MP, si bien el primero **calcula *L* directamente de una matriz de distancias pareada**, mientras que el segundo calcula *L* en base al ajuste entre caracteres discretos y topologías

• Al igual que para los caracteres discretos, encontrar el árbol de distancias óptimo es computacionalmente difícil. Para números chicos de secs. se pueden usar métodos exactos; para números grandes, se emplean métodos heurísticos (aproximados):

1.- **método de los vecinos**
2.- **método de unión de vecinos (NJ)**
3.- **UPGMA**

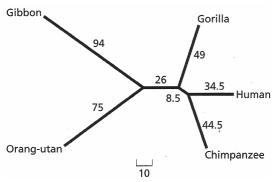
Inferencia filogenética molecular – métodos basados en matrices de distancias

• Criterio de optimización de **Evolución Mínima**

Se pueden encontrar árboles de EM mediante técnicas de **programación lineal** (encontrar una solución óptima dadas unas constricciones). Aplicado a encontrar la longitud de un árbol las constricciones son: 1) ramas de long. = 0; 2) que para cada par de secuencias las distancias topológicas nunca sean < que las observadas ($p_{ij} = d_{ij}$ para todos los pares *ij*)

	Human	Chimp	Gorilla	Orang-utan	Gibbon
Human	–	79	92	144	162
Chimp	79	–	95	154	169
Gorilla	92	**102**	–	150	169
Orang-utan	144	154	150	–	169
Gibbon	**163**	**173**	169	169	–

distancias observadas (*p*) sobre diagonal; distancias topológica bajo la diagonal obtenidas mediante programación lineal



árbol de EM con las long. de ramas calculadas de las dist. observadas *p* usando progr. lineal. La long. total del árbol es 331.5

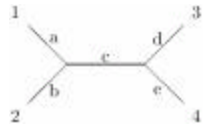
Inferencia filogenética molecular – métodos basados en matrices de distancias

• Criterio de optimización de **Evolución Mínima**

La optimización de long. de ramas mediante PL es computacionalmente costosa para muchos OTUs (>20).

Se usa más frecuentemente el método de mínimos cuadrados para estimar las longitudes de rama. Las long. de rama obtenidas por MC se suman para obtener la *L*

$$SS = \sum_{i<j} \frac{(d_{ij} - p_{ij})^2}{d_{ij}^k}$$



El **método de los mínimos cuadrados** permite encontrar la combinación de valores de (*a*, b, *c*, *d* y *e*) que maximiza el ajuste entre p_{ij} y d_{ij}. Encontrar las long. de ramas mejor ajustadas implica minimizar la suma ponderada de cuadrados.

$w = 1/d K_{ij}$ representa un **factor de ponderación** inversamente proporcional a la distancia estimada, donde $k = 0$ ó $k = 2$. Así las divergencias profundas tienen menor peso que las más recientes, las cuales se pueden estimar mejor.

Inferencia filogenética molecular – métodos basados en matrices de distancias

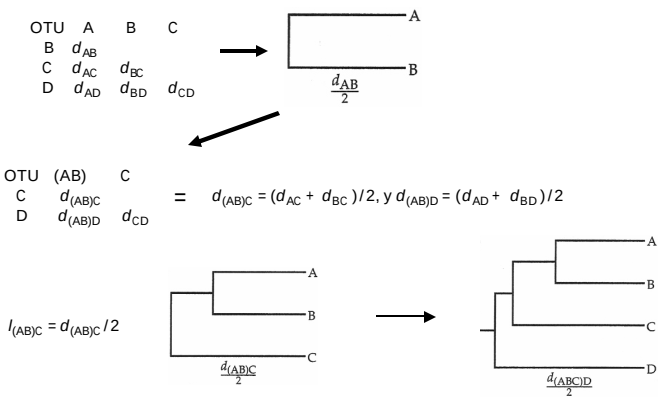
• Unweighted pair group method with arithmetic means (**UPGMA**)

- este es uno de los pocos métodos que construye árboles ultramétricos (todas las hojas equidistantes de la raíz), es decir asume un reloj molecular perfecto a lo largo de toda la topología

- se puede concebir como un método heurístico para encontrar la topología ultramétrica de mínimos cuadrados para una matriz de distancias pareadas

Inferencia filogenética molecular – métodos basados en matrices de distancias

• Unweighted pair group method with arithmetic means (**UPGMA**)



Inferencia filogenética molecular – métodos basados en matrices de distancias

• Unweighted pair group method with arithmetic means (**UPGMA**)

• el punto de ramificación (PR) entre dos OTUs sencillos, *i* y *j*, se posiciona en el punto medio entre ellos

$$l_{ij} = \frac{d_{ij}}{2}$$

• el PR entre un OTU sencillo y uno compuesto (*jm*), se posiciona en el punto medio de la media aritmética de la distancia entre *i* y los constituyentes del OTU compuesto (*jm*)

$$l_{(i)(jm)} = \frac{(d_{ij} + d_{im})/2}{2}$$

• el PR entre dos OTUs compuestos se posiciona a la mitad de la media aritmética de las distancias entre los constituyentes de los OTUs sencillos de cada OTU compuesto. Así el PR entre (*ij*) y (*mn*) es:

$$l_{(ij)(mn)} = \frac{(d_{im} + d_{in} + d_{jm} + d_{jn})/4}{2}$$

• UPGMA, por construir un **árbol ultramétrico**, resulta en una **topología enraizada**
Además se obtienen las longitudes de rama simultáneamente con la topología

Ejercicios del examen de la sección de inferencia filogenética BGE- IV 2005

B) Calcula una matriz de distancias pareadas en base al número observado de diferencias entre OTUs, y en base a ella dibuja un árbol de UPGMA, indicando las longitudes de cada rama

1. Alineamiento: No. sitios : 15; OTUs (taxa) = 4

Rhizobium	GGA GGG AGG AGG CCT
Agrobacterium	GGC GGG AGG AGG CCT
Sinorhizobium	GGG GGA AGG TGT CCG
Bradyrhizobium	GGT CGT AGC TGT GTG

2. Matriz de distancias: *d* : distancia (no. de diferencias observadas)

[	A	B	C	D]
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

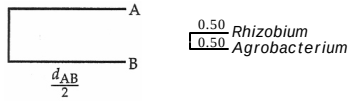
Inferencia de un árbol UPGMA usando el no. de dif. obs. como medida de la distancia genética entre OTUs

Matriz de distancias:

[	A	B	C	D]
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

1.

OTU A B C
B d_{AB}
C d_{AC} d_{BC}
D d_{AD} d_{BD} d_{CD}



0.50 *Rhizobium*
0.50 *Agrobacterium*

2.

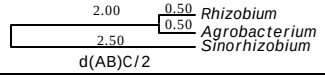
OTU (AB) C
C $d_{(AB)C}$
D $d_{(AB)D}$ d_{CD}

$= d_{(AB)C} = (d_{AC} + d_{BC})/2$, y $d_{(AB)D} = (d_{AD} + d_{BD})/2$

$= d_{(AB)C} = (5 + 5)/2$, y $d_{(AB)D} = (9 + 9)/2$

3.

OTU (AB) C
C 5
D 9 6



Inferencia de un árbol UPGMA usando el no. de dif. obs. como medida de la distancia genética entre OTUs

Matriz de distancias:

[	A	B	C	D]
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

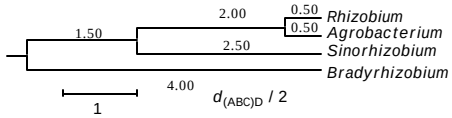
4.

OTU (ABC) D
D $d_{(ABC)D}$

$= d_{(ABC)D} = (d_{AD} + d_{BD} + d_{CD})/3$

$= d_{(ABC)C} = (9 + 9 + 6)/3 = 8$

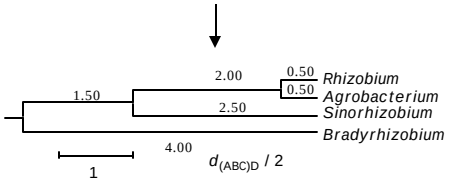
5.



Inferencia de un árbol UPGMA usando el no. de dif. obs. como medida de la distancia genética entre OTUs

Matriz de distancias:

[	A	B	C	D]
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

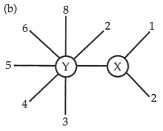


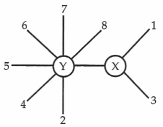
- ¿Notan alguna inconsistencia entre las distancias topológicas y observadas?

- La distancia entre C y D no es aditiva y no queda adecuadamente reflejada en la correspondiente longitud de rama

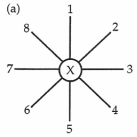
Inferencia filogenética molecular – métodos basados en matrices de distancias

- Método neighbor-joining (**N J**)
- Se trata de un método puramente algorítmico, representando una buena aproximación heurística para encontrar el árbol de evolución mínima más corto. Secuencialmente encuentra vecinos que minimizan la longitud total del árbol
- Es muy rápido y proporciona un solo árbol





N(N-1)/2 modos de buscar pares de OTUs en X



árbol estrella para N OTUS

$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (d_{1k} + d_{2k}) + \frac{1}{2} d_{12} + \frac{1}{N-2} \sum_{3 \le i < j \le N}^{N} d_{ij}$$

- expresión para la suma de todas las long. de ramas
- se busca el par que minimiza S y se considera como un OTU compuesto
- se calcula una nueva matriz de dist. como en UPGMA
- se reitera hasta encontrar todas las N-3 ramas internas

Cálculo de límites de confianza para topologías

- **Exactitud y precisión en filogenética**

En filogenética la **exactitud** de una topología indica su grado de proximidad a la realidad (filogenia verdadera a estimar), mientras que la **precisión** tiene que ver con la cantidad de árboles alternativos que el método es capaz de desechar.

Si tenemos dos termómetros (A y B) con los que medimos la temperatura de agua hirviendo (a 1atm. de presión) y obtenemos las medidas A = 101°C y B = 97.35°C diríamos que A es más exacto pero menos preciso.

Ojo, métodos filogenéticos basados en criterios de optimización que producen puntajes (scores) como números reales como el de máxima verosimilitud (-ln*L* = 3598.2483) dan una falsa impresión de mayor precisión que aquellos CO como el de máxima parsimonia que trabajan con números enteros (no. de pasos mutacionales) (*L* = 257 pasos). En el segundo caso simplemente existe un número finito de pasos mutacionales que definen a las longitudes de un árbol. De ahí que independientemente del método de reconstrucción utilizado para recuperar una filogenia, la precisión de ésta se mide en base al número de topologías alternas que se descartan. Idealmente todas salvo una.

Homoplasias y error de muestreo

- La calidad de la señal filogenética de los datos es una de las fuentes de posible error en la estima filogenética, pudiendo afectar tanto a la exactitud como a la precisión de la estima.
- Si un set de datos contiene **homoplasias** implica que distintos sitios del alineamiento van a apoyar diferentes topologías. Por lo tanto, **qué árbol (o árboles) van a ser apoyados por un set de datos dependerá del subconjunto de caracteres muestreados**.

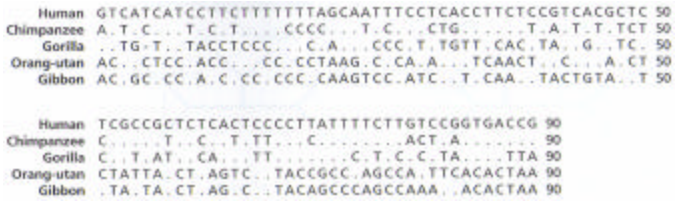
```
Human       GTCATCATCCTTCTTTTTTTAGCAATTTCCTCACCTTCTCCGTCACGCTC 50
Chimpanzee  A.T.C...T.C.T....CCCC...T.C...CTG.....T.A.T.T.TCT 50
Gorilla     ..TG.T..TACCTCCC...C.A...CCC.T.TGTT.CAC.TA..G..TC. 50
Orang-utan  AC..CTCC.ACC...CC.CCTAAG.C.CA.A...TCAACT..C...A.CT 50
Gibbon      AC.GC.CC.A.C.CC.CCC.CAAGTCC.ATC..T.CAA..TACTGTA..T 50

Human       TCGCCGCTCTCACTCCCCTTATTTTCTTGTCCGGTGACCG 90
Chimpanzee  C.....T..C..T.TT...C........ACT.A..... 90
Gorilla     C..T.AT..CA...TT.......C.T.C.C.TA....TTA 90
Orang-utan  CTATTA.CT.AGTC..TACCGCC.AGCCA.TTCACACTAA 90
Gibbon      .TA.TA.CT.AG.C..TACAGCCCAGCCAAA..ACACTAA 90
```

90 sitios parsimonia informativos (de 986 sitios de *coi*) que resultan en el árbol:
((human,(chimp,gorilla)),orang,gibbon).

Homoplasias y error de muestreo

Pero si se muestrean sólo los primeros 31 sitios del aln (5 sitios Pi) obtendríamos un árbol de MP con la siguiente topología: (((human,gorilla),chimp),orang,gibbon), que no se corresponde con el árbol de MP para el set completo de datos. El primer sito apoya (human, gorilla), el 2° (human,chimp,gorilla) y la 3°. (chimp,gorilla), que contradice a la relación apoyada por la 1a. pos.

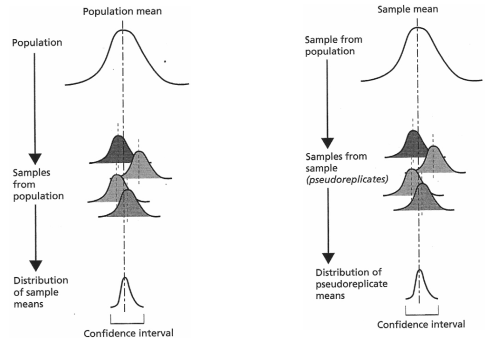


El muestreo de las aprox. 16.000 pb del genoma mitocondrial de estos primates y sets de datos más extensos (con más OTUs) soportan el árbol:
(((human,chimp),gorilla),orang,gibbon).

Por tanto, para minimizar los errores de muestreo (debidos a homoplasias) hay que tratar de obtener secuencias lo más largas posibles para el mayor número posible de genes

Estima del error de muestreo mediante el método de bootstrap

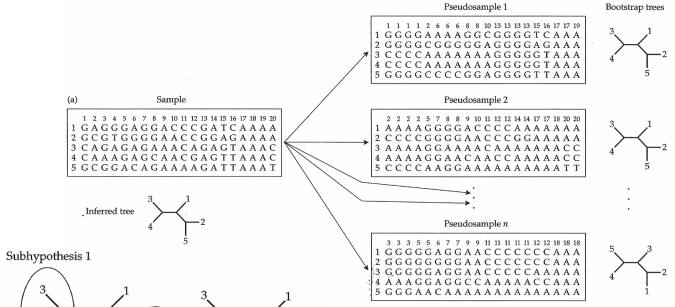
- Una vía de estimar el error de muestreo es tomar múltiples muestras de la población y comparar las estimas obtenidas de ellas. La dispersión entre estas muestras nos da una idea del error de muestreo
- El **método de bootstrap se basa en remuestrear una muestra única**



Estima del error de muestreo en un alineamiento múltiple mediante el método de bootstrap

Alineamiento original →

1. Pseudoréplicas de bootstrap (muestreo aleatorio de caracteres con reemplazo hasta obtener una matriz de igual no. de caracteres que la original

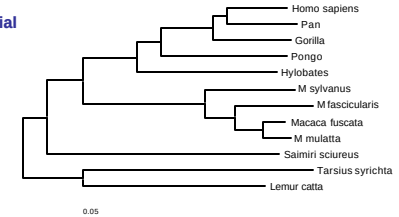


2. Obtener el árbol consenso que resume la información de las topologías recuperadas de cada pseudoréplica de bootstrap.

3. Mapear las proporciones de bootstrap sobre la topología original

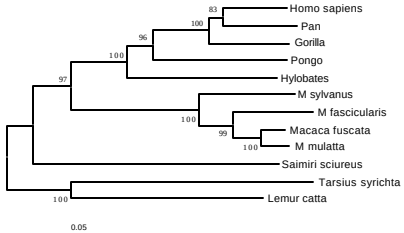
Estima del error de muestreo mediante el método de bootstrap

Filogenia NJ-K2P+G estimada a partir de secuencias de **NADH-DH mitocondrial** de 12 primates.



¿Están todas las bipartciones igualmente soportadas? O planteado de otra manera: ¿qué tan resuelta está esta filogenia?

Un **análisis de bootstrap con 500 pseudoréplicas** indica que el árbol está muy bien resuelto:



todas las bipartciones están áltamente justificadas por los datos según indican las proporciones de bootstrap (expresadas en porcentaje)