Introducción a la Inferencia Filogenética y Evolución Molecular

23-26 Junio 2008, Fac. C. Biológicas - UANL

Pablo Vinuesa (vinuesa@ccg.unam.mx)

Centro de Ciencias Genómicas-UNAM, México
http://www.ccg.unam.mx/~vinuesa/

Todo el material del curso lo puedes descargar desde:

http://www.ccg.unam.mx/~vinuesa/UANL08

- **Tema 7: Métodos de reconstrucción basados en matrices de distancias:**

1. Caracteres vs. distancias, distancias genéticas vs. topológicas
2. Criterios de optimización usados con matrices de distancias: Mínimos cuadrados y Evolución mínima
3. Métodos algorítmicos: UPGMA y neighbor-joining
4. Prácticas usando los programas BioEdit, DAMBE y MEGA4 (windows)

Inferencia filogenética molecular – Métodos de distancia

- **Tipos de datos:**
  - **caracteres**: proveen información sobre cada 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. generalmente contínuos; 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:**

- **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?**

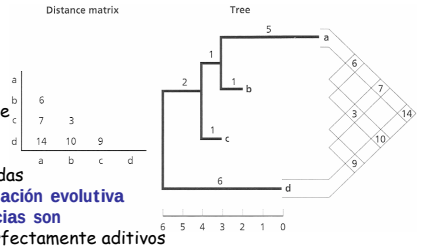
1.- Una larga lista de estados de caracter, como una secuencia de DNA ó aa, carece en sí misma de significado evolutivo; en cambio, decir que 3 secuencias A <-> B <-> C presentan 95% y 50% de identidad entre ellas evoca una imagen intuitiva del "grado de parentesco"

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)

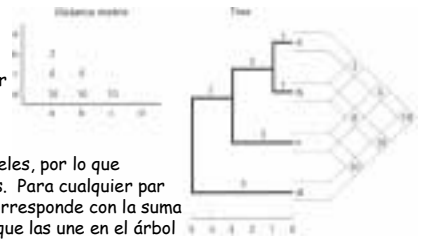
3.- Los métodos de reconstruccón basados en matrices de dist. son muy rápidos

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

Métodos basados en matrices de distancias – Criterios de optimización

- 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:

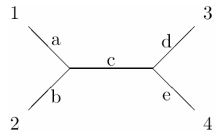
1. **métodos de "bondad de ajuste":** buscan el árbol métrico que mejor acomoda las distancias "observadas" usando el método de **mínimos cuadrados**

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

Métodos basados en matrices de distancias – Criterios de optimización

- **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

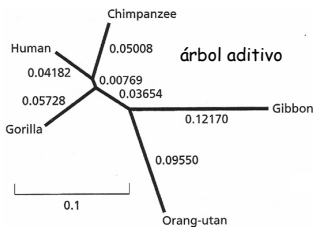
diag. infer.: dist. evolutivas

Métodos basados en matrices de distancias – Criterios de optimización

- **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 MC para mtDNAs.
En **negritas** ***dt > de***, en *cursiva dt < de* (*dt* =dist. topol.; *de* = dist. observada o evolutiva)



- Las ***dt > de*** pueden explicarse por homoplasias en algunas ramas
- Las *dt < de* no pueden explicarse fácilmente y son contra-intuitivas, ya que implicarían que aconteció menos cambio evolutivo que el observado!
- Ello ha llevado a algunos investigadores a criticar fuertemente el método de los MC para estimar la long. de las ramas

Métodos basados en matrices de distancias – Criterios de optimización

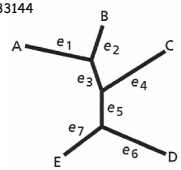
- Método de los mínimos cuadrados (medidas de la "bondad de ajuste")

(2n-3) = 7 ramas independientes

(n-1) = 4 ramas independientes

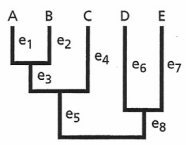
Additive tree

SS = 0.000033144



Ultrametric tree

SS = 0.26577



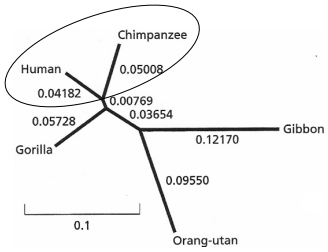
Constraints

$e_1 = e_2$
$e_6 = e_7$
$e_4 = e_1 + e_3$
$e_4 + e_5 = e_6 + e_8$

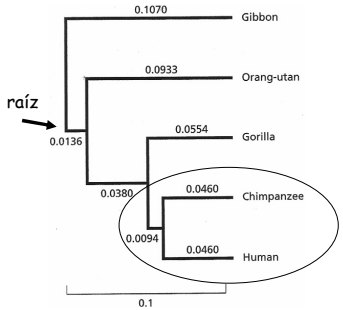
- topologías aditivas y ultramétricas para las mismas secuencias. La topología ultramétrica tiene menor número de ramas con longitudes únicas dadas las restricciones impuestas
- A mayor desvío del reloj molecular (igualdad de tasas evolutivas entre linajes) mayor desvío de la ultrametricidad de los datos y mayor la diferencia en el ajuste entre los árboles aditivos y ultramétricos a los datos
- Los aditivos tienen entonces mayor ajuste dado que no tienen restricciones de ultrametric.

Métodos basados en matrices de distancias – Criterios de optimización

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



árbol aditivo



árbol ultramétrico

Métodos basados en matrices de distancias – Criterios de optimización

- **Criterio de Evolución Mínima**

- dada una **topología aditiva** para *n* secuencias, existen **(2*n* - 3) ramas**, cada una con una longitud *l*. 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**

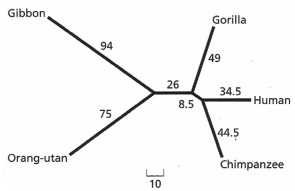
Métodos basados en matrices de distancias – Criterios de optimización

- **Criterio 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 ($p_{ij} \geq d_{ij}$ 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

Métodos basados en matrices de distancias – Criterios de optimización

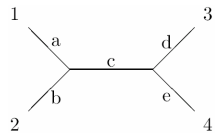
- Criterio de **Evolución Mínima**

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.

Métodos basados en matrices de distancias – Métodos algorítmicos

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

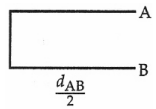
- este es uno de los pocos métodos que construye **árboles ultramétricos** (todas las hojas equidistantes de la raíz), es decir **asume un reloj molecular** perfecto a lo largo de toda la topología, lo que resulta en una **topología enraizada**.
  Además se obtienen las longitudes de rama simultáneamente con la topología

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

Métodos basados en matrices de distancias – Métodos algorítmicos

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

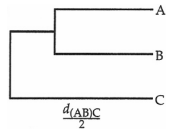
OTU	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

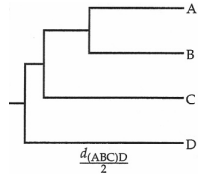


OTU	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

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

$l_{(AB)C} = d_{(AB)C}/2$





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

Ejercicio:

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	

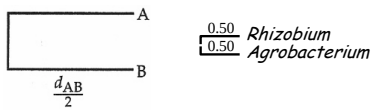
Inferencia de un árbol UPGMA usando el no. de dif. obs. como medida de la distancia genética entre OTUs

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}



2.

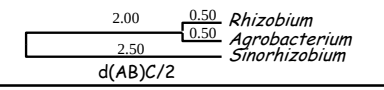
OTU	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

= $d_{(AB)C} = (d_{AC} + d_{BC})/2$, 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:

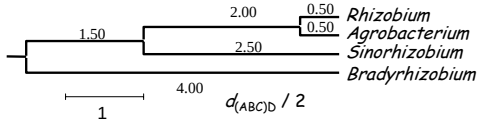
	A	B	C	D]
[*Rhizobium*, A]				
[*Agrobacterium*, B]	1.0			
[*Sinorhizobium*, C]	5.0	5.0		
[*Bradyrhizobium*, D]	9.0	9.0	6.0	

4. OTU (ABC) D; D $d_{(ABC)D}$

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

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

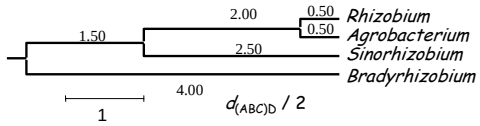
5.



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	

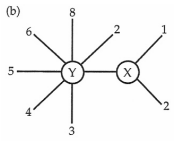


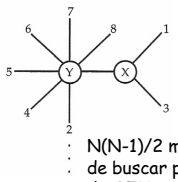
- ¿Notan alguna inconsistencia entre las distancias topológicas y observadas?

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

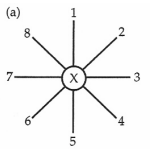
Métodos basados en matrices de distancias – Métodos algorítmicos

- Método neighbor-joining (**NJ**)
- 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 aditivo (no ultramétrico).





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