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 6: Modelos de evolución de secuencias y prueba de bootstrap

1. El uso de modelos en ciencia y en filogenética
2. Modelos empíricos vs. paramétricos
3. Derivación de matrices de sustitución empíricas a partir de alineamientos múltiples de proteínas
4. Modelos paramétricos de evolución de secuencias de DNA – la familia GTR
5. Modelos y corrección de distancias genéticas
6. La prueba de bootstrap para determinar la confiabilidad estadística de biparticiones

Modelos de evolución de secuencias -introducción

· **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

1.- La reconstrucción o estima filogenética es un **problema de inferencia estadística**, y como tal **requiere un modelo** de sustitución de resíduos (aa o nt), es decir, un modelo de evolución molecular de las secuencias. Todos los modelos, por no ser más que aproximaciones de los procesos naturales, hacen una serie de **suposiciones** (simplificaciones)

2.- Los **modelos de evolución de secs.** son usados en filogenética **para describir las probabilidades con las que se dan los distintos eventos de sustitución entre aa o nt**, con el fin de **corregir o compensar las sustituciones no observadas a lo largo de la filogenia**

3.- Mientras que los métodos de MP asumen un **modelo implícito** de evolución (número mínimo de sustituciones a lo largo de la filogenia), los métodos de distancia (UPGMA, NJ), los de ML y Bayesianos requieren de un **modelo explícito** de evolución

4.- Los **métodos de distancia** estiman finalmente un sólo parámetro (no. sust./sitio) dado el modelo y el valor de los parámetros del mismo; en cambio, los **métodos de ML y Bayesianos pueden estimar el valor de cada uno de los parámetros del modelo explicitado**, dada una topología y la matriz de datos (alineamiento)

Modelos de evolución de secuencias -introducción

· Para el análisis filogenético de secuencias alineadas virtualmente todos los métodos describen la evolución de las secuencias usando un modelo que consta de dos componentes:

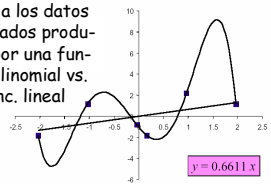
1. un árbol filogenético
2. una descripción de las probabilidades con las que se dan las sustituciones de aa o nts a lo largo de las ramas del árbol

· ¿Porqué necesitamos modelos y para qué sirven?

- Los modelos nos sirven para interpolar adecuadamente entre nuestras observaciones con el fin de poder hacer predicciones inteligentes sobre observaciones futuras

$y = -1.5972\,x^5 + 23.167\,x^4 - 126.18\,x^3 + 319.17\,x^2 - 369.22\,x + 155.67$

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

Modelos de evolución de secuencias -introducción

· **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

Corolario:

1. El grado de confianza que tengamos en una filogenia particular realmente depende de la que tengamos en el modelo subyacente

2. Por lo tanto, siempre que usemos un método basado en un modelo explícito de evolución (NJ, ML, By) es necesario usar rigurosas pruebas estadísticas para seleccionar el modelo y el valor de sus parámetros que mejor se ajusten a la matriz de datos a analizar

Modelos de evolución de secuencias -introducción

- **Modelos de evolución del proceso de sustitución y métodos de reconstrucción filogenética: consideraciones generales**

- Existen dos aproximaciones para construir modelos de evolución de secuencias.

1. construcción de **modelos empíricos** basados en propiedades del proceso de sustitución calculadas a partir de comparaciones de un gran número de secuencias. Los modelos empíricos **resultan en valores fijos de los parámetros**, los cuales son estimados sólo una vez, suponiéndose que son adecuados para el análisis de otros sets de datos. Esto los hace fácil de usar e implementar en términos computacionales, pero su utilidad real para cada caso particular ha de ser evaluada críticamente

2. construcción de **modelos paramétricos** basado en el modelaje de propiedades químicas o genéticas del aas y nts. Los modelos paramétricos tienen la ventaja de que los **valores de los parámetros pueden ser derivados de cada set de datos** al hacer un análisis de los mismos usando métodos de ML o By, por tanto ajustándolos a cada matriz de datos particular

Similitud entre pares de secuencias de AA

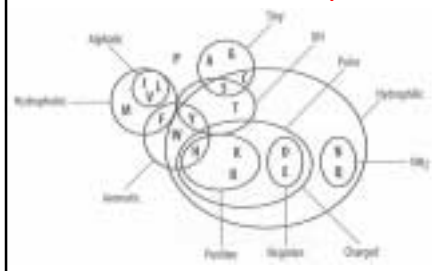


Table 2 - The log odds matrix for BLOSUM 62

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C		9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-1	-2
D			6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E				5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F					6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G						6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H							8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I								4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K									5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L										4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M											5	-2	-2	0	-1	-1	-1	1	-1	-1
N												6	-2	0	0	1	0	-3	-4	-2
P													7	-1	-2	-1	-1	-2	-4	-3
Q														5	1	0	-1	-2	-2	-1
R															5	-1	-1	-3	-3	-2
S																4	1	-2	-3	-2
T																	5	0	-2	-2
V																		4	-3	-1
W																			11	2
Y																				7

Matriz BLOSUM62

- **Las matrices empíricas de sustitución** entre AAs no reflejan necesariamente las relaciones químicas entre ellos. Se trata de una **definición púramente estadística basada en el análisis de frecuencias empíricas de sustituciones observadas en alineamientos de secs. con un grado de divergencia definido**

- Cada score de la matriz representa la tasa de sustitución esperada entre un par de AAs. Por tanto, los scores de los alineamientos pareados evaluados con estas matrices reflejan la distancia evolutiva existente entre las secuencias.
Es importante notar que los scores son evolutivamente simétricos al no conocerse la dirección del cambio evolutivo.

Alineamiento pareado de proteínas: matrices de costo BLOSUM

Matrices BLOSUM de sustitución de aa
Henikoff, S., Henikoff, J. G., and Pietrokovski, S. 1999. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics **15**: 471-479.

- Desarrollada por S. Henikoff y J. G. Henikoff para obtener una matriz más robusta que las PAM en la identificación de homólogos distantes, particularmente cuando contienen una proporción significativa de aas hidrofóbicos

- Las matrices **BLOSUM** están basadas en la base de datos **BLOCKS+** de proteínas alineadas; **BLO**cks **SU**bstitution **M**atrix (http://blocks.fhcrc.org). Son matrices empíricas.

- Las series de matrices **BLOSUM** se derivaron de alineamientos sin indeles **(BLOCKS)** de proteínas **considerando sólo pares de alineamientos que no divergieran más de un umbral determinado**, por ej. un **mínimo de 62 % de identidad**, para calcular las frecuencias diana o esperadas de la **matriz BLOSUM62**. Para estos alns. se calcula la razón entre el número de pares de aa observados en cada posición y el número de pares esperados de las frequencias globales de los aas, expresando los resultados como $\log_{10} X \lambda$

- Para evitar sesgos en las matrices por sobrerepresentación de secuencias muy similares, se reemplazaron aquellas con similitud > a un umbral dado por un solo representante o por un promedio ponderado (BLOCKS+).

- **La matriz BLOSUM62 es la actualmente favorecida para la mayoría de las aplicaciones por su buen rendimiento empírico y ha reemplazado a las matrices de Dayhoff (PAM)**

Alineamiento de proteínas: selección de matrices de ponderación - consejos prácticos

- Las **matrices PAM** fueron derivadas de las secuencias de proteínas disponibles a finales de los 60s y ppios. de los 70s. Era una base de datos muy reducida y estaba **sesgada a proteínas chicas, globulares e hidrofílicas**! Al carecer de suficientes homólogos con diversos niveles de divergencia evolutiva tuvieron que emplear supuestos teóricos (extrapolación) para obtener las matrices de sustitución para prots. más distantes (mediante exponenciación)

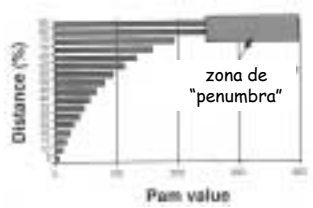
- las matrices PAM son una pobre elección para alinear (o buscar en las bases de datos) proteínas con dominios hidrofóbicos (p. ej. dominios transmembrana)

- Qué matriz escoger en función del nivel de divergencia esperada **(potencial de mira retrospectiva en tiempo evolutivo)**

% identidad	PAM	BLOSUM	mira retrospectiva en tiempo evolutivo
20- 50 %	250	45	homólogos en la zona de penumbra
50- 75 %	250	62	ortólogos y parálogos en superfamilias[1]
75- 90 %	160	80	ortólogos y parálogos en familias[2]
90- 99 %	40	90	ortólogos muy cercanos

[1]Superfamilias de proteínas contienen diversas familias de proteínas con ≥ 30% identidad entre ellas
[2]Familias de proteínas contienen secuencias con ≥ 85% identidad entre ellas
Estas definiciones fueron acuñadas por Dayhoff et al. (1978)

Alineamiento de proteínas: selección de matrices de ponderación - consejos prácticos para la identificación de homólogos



Distancias observadas vs. evolutivas (PAM) entre prots.

Diferencia % obs.	Dist. evol. PAM	
1	1	
5	5	
10	11	
15	17	
20	23	
30	38	
40	56	
50	80	
60	112	
70	159	
80	246	← z. penumbra
85	328	

- A medida que el nivel de divergencia entre pares de proteínas alcanza el valor de PAM250 (~ **20% identidad**), comienza a ser dudosa su relación de homología, pudiendo tratarse de secuencias que presentan cierto grado de similitud por azar, en base a composiciones de AAs similares en ambas secuencias !!!
- Al entrar en esta **zona de penumbra**, es esencial considerar información adicional, particularmente motivos estructurales, para validar o descartar una posible relación de homología
- A medida que el nivel de divergencia evolutiva entre pares de proteínas incrementa (distancias PAM) disminuye el número de diferencias observadas, debido a fenómenos de **reversión** (**homoplasia**). Por tanto, si no se cuenta con evidencia estructural, el análisis filogenético de proteínas debe restringirse a aquellas con ≥ 20% de identidad. **Los alns. tampoco son confiables**

Modelos de evolución de secuencias -DNA

· Modelos de sustitución de nucleótidos

- El modelaje de la evolución a nivel del DNA se ha concentrado en la aproximación paramétrica. Se manejan **tres tipos principales de parámetros** en estos modelos:

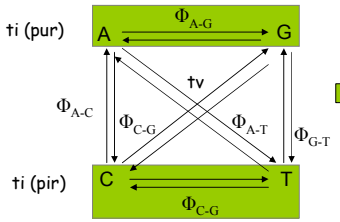
1. parámetros de **frecuencia**
2. parámetros de **tasas de intercambio**
3. parámetros de **heterogeneidad de tasas de sustitución** entre sitios

Modelos de evolución de sustitución de nucleótidos -modelos paramétricos

- **los diversos modelos evolutivos se distinguen por su grado de parametrización**

I. Frecuencias de nt : $\pi A = \pi C = \pi G = \pi T = 0.25$ ó $\pi A \neq \pi C \neq \pi G \neq \pi T$
- modelos de = frecuencia: JC69; K2P, K3P ...
- modelos de ≠ frecuencia: F81, HKY85, TrN93, GTR ...

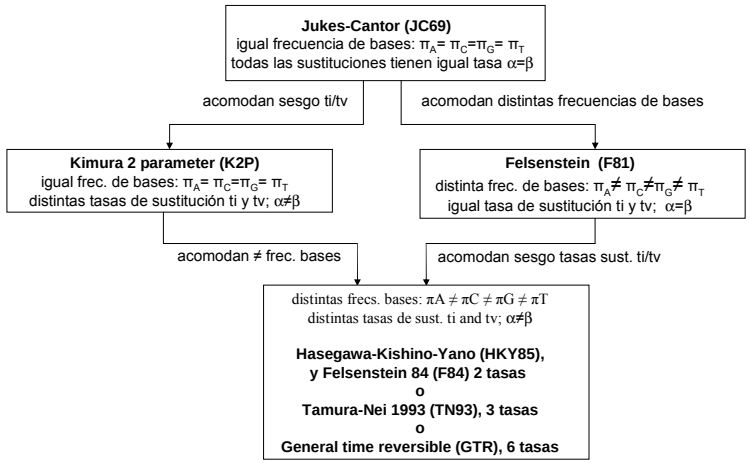
II. Tasas de sustitución transicionales/transversionales



- Existen 4 tipos de sustituciones ti y 8 tv; cuando **ti/tv ≠ 0.5 existe un sesgo en sustituciones ti (o tv) en el set de datos. ti generalmente >> 1**
- los modelos evolutivos se diferencian también en la cantidad de parámetros que utilizan para acomodar diversas tasas de sustitución:

tasas	modelo
1	**JC69** (ti=tv)
2	**K2P** (ti ≠tv)
3	**TrN** ó **K3P** (2 ti, 1 tv)
6	**GTR** (cada sust. su tasa)

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV



Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV

- Matriz de tasas de sustitución instantáneas del modelo GTR

	A	C	G	T
A	–	$\pi_C a\mu$	$\pi_G b\mu$	$\pi_T c\mu$
C	$\pi_A a\mu$	–	$\pi_G d\mu$	$\pi_T e\mu$
G	$\pi_A b\mu$	$\pi_C d\mu$	–	$\pi_T f\mu$
T	$\pi_A c\mu$	$\pi_C e\mu$	$\pi_G f\mu$	–

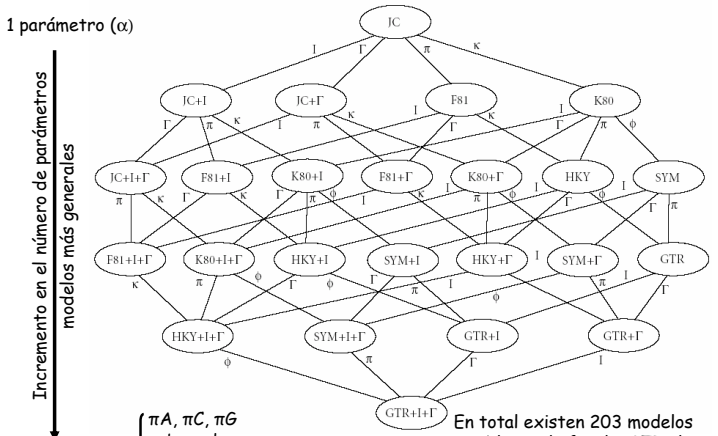
$-\mu(\pi_A c + \pi_C e + \pi_G f)$

9 parámetros: π_A, π_C, π_G, a, b, c, d, e, μ

El modelo GTR es idéntico al de JC69 si $a = b = c = d = e = f = 1$ y todas las bases se asumen que tienen igual frecuencia ($\frac{1}{4}$)

μ = tasa del proceso generador de todos los tipos de sustituciones
$a, \ldots e$ = modificadores de tasa relativa de cada tipo particular de sustitución
π = frecuencia de cada nt

Modelos básicos de evolución de DNA: la familia de modelos anidados GTR o REV



11 parámetros libres a estimar: πA, πC, πG, a, b, c, d, e, μ, I, T

En total existen 203 modelos posibles en la familia GTR al combinar params. de frec., tasa, G e I
La mayoría de ellos carecen de nombre

Comparación empírica de modelos sust. de DNA

- Comparación de los modelos de **JC69 y K2P** en su capacidad de corregir distancias observadas (p) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

- Escenario I:

  - sean 2 secs. de long. = 200 nt, que difieren en 20 *ti* y 4 *tv*

  por lo tanto L = 200, P = 20/200 = 0.1 y Q = 4/200 = 0.02

p = 24/200 = 0.12

$d_{JC69} \approx 0.13$ (sust./sitio)

no. de sust. esperadas = 0.13 X 200 ≈ 26

$d_{K2P} \approx 0.13$ (sust./sitio)

no. de sust. esperadas = 0.13 X 200 ≈ 26

Comparación empírica de modelos sust. de DNA

- Comparación de los modelos de **JC69 y K2P** en su capacidad de corregir distancias observadas (p) entre pares de secuencias según su grado de divergencia

$$d_{JC69} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \quad \text{vs.} \quad d_{K2P} = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

- Escenario II:

  - sean 2 secs. de long. = 200 nt, que difieren en 50 *ti* y 16 *tv*

  por lo tanto L = 200, P = 50/200 = 0.25 y Q = 16/200 = 0.08

p = 66/200 = 0.33

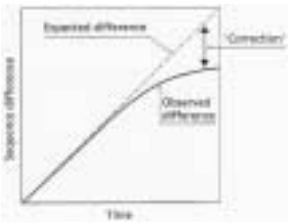
$d_{JC69} \approx 0.43$ (sust./sitio)

no. de sust. esperadas = 0.43 X 200 ≈ 86

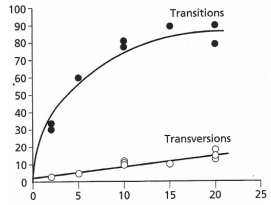
$d_{K2P} \approx 0.48$ (sust./sitio)

no. de sust. esperadas = 0.48 X 200 ≈ 96

Modelos de evolución de secuencias



- El objetivo de los modelos de sustitución es el de **compensar para los eventos homoplásicos de múltiples sustituciones**, y así obtener estimas de distancias evolutivas corregidas



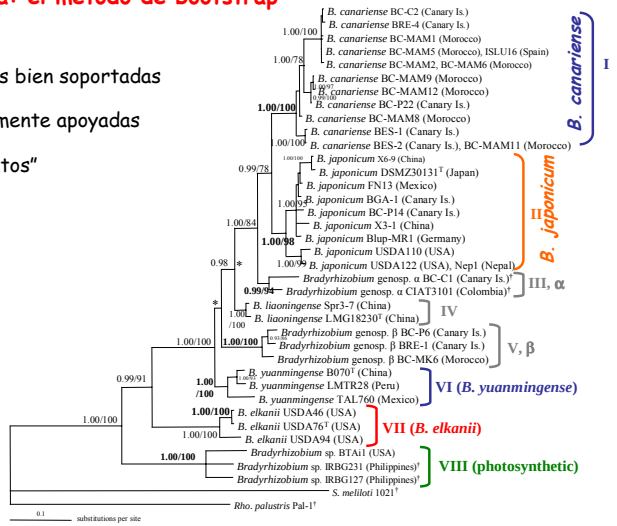
- El número de *ti*'es generalmente > que el de *tv*, fenómeno que se acentúa cuanto mayor es la divergencia entre las secuencias a comparar. De ahí que en nuestro ejemplo las diferencias entre los escenarios I y II sólo se hicieron notar en el caso en el que la divergencia entre las secuencias era mayor (escenario II)

Estima de la confianza que podemos tener en distintas partes de una filogenia: el método de bootstrap

"Filogenias bien soportadas

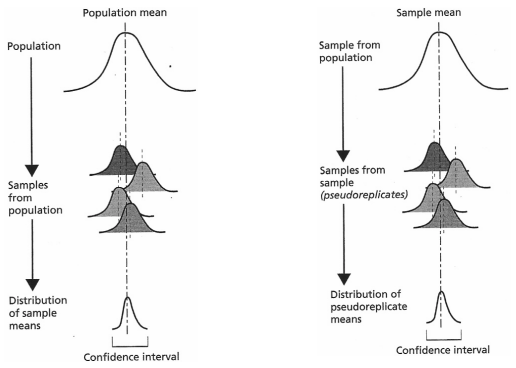
vs. pobremente apoyadas

por los datos"

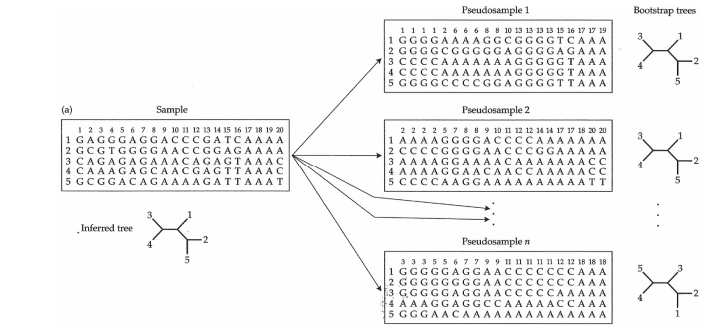


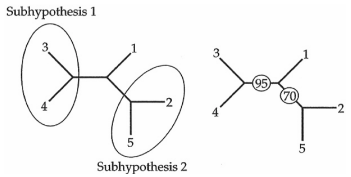
Estima del error de muestreo mediante el método de bootstrap

- 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 la propia muestra**



Estima del error de muestreo mediante el método de bootstrap





BOOTSTRAPPING

- generación de n (100-1000) **pseudoréplicas** (muestreo aleatorio con reemplazo)
2. estima de la filogenia para cada pseudo-réplica
3. cálculo de un **árbol consenso**
4. mapeo de las proporciones de bootstrap sobre la topología inferida de los datos originales