

Caracteres moleculares

Esquema para esta clase...

- Generalidades
- Tipos de caracteres moleculares
- Particularidades del análisis de caracteres moleculares (en referencia al análisis de secuencias)
- Caracteres moleculares y morfológicos
- Consideraciones finales

Generalidades

Mensaje:

Conceptualmente un carácter molecular es lo mismo que un carácter morfológico

(son atributos heredables de los organismos)

Por lo tanto: los caracteres morfológicos y los moleculares pueden analizarse de la misma forma

Algunas razones de la popularidad de la sistemática molecular:

- estrecha relación con la **Evolución Molecular**
- ha **rejuvenecido** a la sistemática
 - ha permitido documentar diversidad que de otra forma pasaba desapercibida
 - propiciado avances conceptuales y refinamientos metodológicos
 - propiciado debates

Tipos de caracteres moleculares

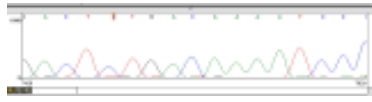
Proteínas

- distancias inmunológicas
- electroforesis
- secuencias

Ácidos Nucleicos (ADN o ARN)

- hibridización de ADN-ADN
- polimorfismos en el largo de fragmentos de ADN
- orden de genes en el genoma (bacteriano o mitocondrial)
- SINES y LINES ("Short and Long Interspersed Retrotransposable Elements")
- microsátélites, SNPs.

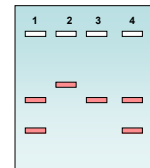
- **secuencias**



Electroforesis de proteínas

- Caracteriza a las proteínas por tamaño y/o carga eléctrica

- Provee caracteres discretos:
presencia/ausencia de una banda
que se pueden transformar a frecuencias



- Casi no se usa en estudios filogenéticos
- Cada vez menor aplicación en estudios poblacionales

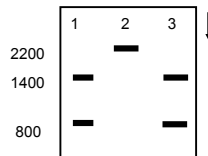
Polimorfismos en el Largo de Fragmentos de Restricción (RFLPs)

Combina la amplificación de ADN y el corte del mismo usando enzimas de restricción

- 1) ...A A A T C A T T C G A A T T A A G G G G T T...
- 2) ...A G G T C A T T C G A A T T A G A G A G T T...
- 3) ...A G G T C A T T C G A A T T A G A G A G T T...

ej: Enzima de restricción X:

- reconoce secuencia **TTCGA**
- y corta después de la 2da T



Pero....

- no revela toda la variación existente a nivel de las secuencias

- 1) ...A A A T C A T T C G A A T T A A G G G G T T...
- 2) ...A G G T C A T T C G A A T T A G A G A G T T...
- 3) ...A G G T C A T T C G A A T T A G A G A G T T...

- uso de una batería de enzimas

igual no revela toda la variación existente

- provee caracteres discretos:
bandas presentes o no
- hoy casi no se usa para reconstruir relaciones filogenéticas
- cada vez menor aplicación en estudios poblacionales

Secuencias de ADN

Secuencias: fragmentos de ADN o genomas completos
(amplificados por PCR)

- ADN mitocondrial
- ADN de cloroplastos
- ADN nuclear
- ARN

información más directa para inferir relaciones filogenéticas

secuencias nucleotídicas = tipo de caracteres moleculares más usados

Análisis de secuencias

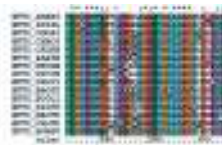
1^{er} paso – “**aline**ar” las secuencias

y esto no es trivial

AACCAATTGG
ACATG

2^{do} paso – reconstrucción filogenética

Un **alineamiento** es una hipótesis sobre las homologías posicionales entre bases (caracteres) de distintos individuos



Es decir: el establecimiento de la correspondencia de bases entre las secuencias de los distintos especímenes

Un **alineamiento** es una hipótesis sobre las **homologías posicionales** entre bases (caracteres) de distintos individuos

```
1) ...G C C T A C C...
2) ...G C C T A C C...
3) ...G C A T A C C...
4) ...G C C T A C C...
5) ...G C C T A C C...
6) ...G A C T A C C...
7) ...G A T A C C...
```

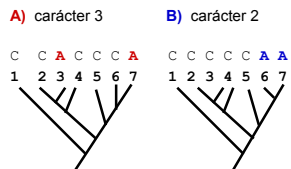
A) 1 2 3 4 5 6 7
1) G C C T A C C
2) G C C T A C C
3) G C A T A C C
4) G C C T A C C
5) G C C T A C C
6) G A C T A C C
7) G - A T A C C

B) 1 2 3 4 5 6 7
1) G C C T A C C
2) G C C T A C C
3) G C A T A C C
4) G C C T A C C
5) G C C T A C C
6) G A C T A C C
7) G A - T A C C

Un **alineamiento** es una hipótesis sobre las **homologías posicionales** entre bases (caracteres) de distintos individuos

A) 1 2 3 4 5 6 7
 1) G C C T A C C
 2) G C T A C C
 3) G C A T A C C
 4) G C C T A C C
 5) G C C T A C C
 6) G A C T A C C
 7) G - A T A C C

B) 1 2 3 4 5 6 7
 1) G C C T A C C
 2) G C T A C C
 3) G C A T A C C
 4) G C C T A C C
 5) G C C T A C C
 6) G A C T A C C
 7) G A - T A C C



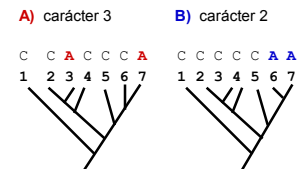
Supongamos que las relaciones entre los 7 taxones están dadas por la siguiente filogenia

Al mapear los estados de los caracteres 3 y 2 en los alineamientos A) y B) respectivamente tenemos:

Un **alineamiento** es una hipótesis sobre las **homologías posicionales** entre bases (caracteres) de distintos individuos

A) 1 2 3 4 5 6 7
 1) G C C T A C C
 2) G C T A C C
 3) G C A T A C C
 4) G C C T A C C
 5) G C C T A C C
 6) G A C T A C C
 7) G - A T A C C

B) 1 2 3 4 5 6 7
 1) G C C T A C C
 2) G C T A C C
 3) G C A T A C C
 4) G C C T A C C
 5) G C C T A C C
 6) G A C T A C C
 7) G A - T A C C



A en 3 y A en 7 son:

estados de carácter análogos (el ancestro común de 3 y 7 no tenía A)

A en 6 y A en 7 son:

estados de carácter homólogos (el ancestro común de 6 y 7 tenía A)

Alineamiento de estas dos secuencias: 1 AACCAATTGG
 2 ACATG

Dos alineamientos (de los muchos posibles):

1 AACCAATTGG
 2 -AC--AT-G- 4 gaps + 0 cambio
 ** ** *

1 AACCAATTGG
 2 ACATG----- 1 gap + 4 cambios
 *

Gaps: representan eventos de **inserción o delección** de bases ("indel")

Cambios: son **sustituciones** de bases

purinas = A(denina) y G(uanina)

pirimidinas = T(imina) y C(itosina)

transiciones (Ts): A ↔ G T ↔ C (4 tipos)

transversiones (Tv): A ↔ T A ↔ C G ↔ T G ↔ C (8 tipos)

Las Ts son más comunes que las Tv

¿Como elegir entre los distintos alineamientos?

¿Que criterio usar?

```

1 AACCAATTGG
2 -AC--AT-G-
  **  ** *

1 AACCAATTGG
2 ACATG-----
  *
    
```

Recordar: los datos determinan los resultados
el alineamiento determina la topología que se obtendrá

Respuesta: Balance entre **costos de gaps (indels)** y de **sustituciones**

gap: apertura / extensión
sustituciones: transiciones / transversiones

Asumiendo los siguientes costos:

Gap de apertura = 1
Gap de extension = 0
Cualquier cambio = 1

elegimos el 1er alineamiento

gap cambio

```

1 AACCAATTGG
2 -AC--AT-G-
  **  ** *
4 + 0 = 4

1 AACCAATTGG
2 ACATG-----
  *
1 + 4 = 5
    
```

Sin embargo...si cambiamos los costos:

(gap de apertura ahora cuesta 2)

costo	gap cambio		gap cambio	
	1	1	2	1
1 AACCAATTGG 2 -AC--AT-G- ** ** *	4	+ 0 = 4	8	+ 0 = 8
1 AACCAATTGG 2 ACATG----- *	1	+ 4 = 5	2	+ 4 = 6

El alineamiento final puede variar según los costos asumidos

al menos se debería explorar el efecto de distintas combinaciones de costos y evaluar la variación en el alineamiento final

Además...

Para un juego de costos puede haber más de un alineamiento igualmente óptimo

Gap de apertura = 1
Gap de extension = 0
Cualquier cambio = 1

gap cambio

gap cambio

```

TTAAGAAC 1 + 1 = 2
TAA--AAC
* * ***

TTAAGAAC 2 + 0 = 2
T-AA-AAC
* ** ***
    
```

Sin embargo:

la mayoría de los programas de alineamiento (e.g., ClustalX) dan un único alineamiento

Recordar: el alineamiento determina la topología que se obtendrá

Notar:

- hemos alineado dos secuencias muy cortas

- método usado: "ojo"

- imposible de realizar con una base de datos real

- operativamente es muy complejo

necesidad de explorar un espacio n dimensional
distintos modos para explorar este espacio



Resumen alineamiento:

Para un juego de secuencias tenemos que:

- el alineamiento final depende de los costos asumidos

- para un juego de costos puede haber más de un alineamiento igualmente óptimo

La topología que se obtendrá depende directamente del alineamiento empleado

Avance conceptual

En última instancia lo que se quiere encontrar es:

el alineamiento que produce el árbol más corto



Entonces: el alineamiento de las secuencias y la búsqueda del árbol son parte del mismo problema

optimización del alineamiento

Optimización del alineamiento

- hay programas que hacen esto en un contexto de máxima parsimonia

la optimización del alineamiento también se podría implementar en un marco de máxima verosimilitud

- proceso sumamente complejo por la cantidad de cálculos que se deben de hacer.

imaginarse esto cuando se tienen 1000-3000 pb para 50-100 individuos

¿Qué caracteres son mejores: los morfológicos o los moleculares?

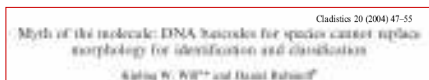
Históricamente los datos moleculares han sido “propagandeados” por algunos sistemáticos moleculares como superiores

Importante notar lo siguiente:

La inmensa mayoría de los estudios filogenéticos (filogeográficos y poblacionales) implican una identificación basada en morfología de los especímenes estudiados

Algunas propuestas de aquellos que sostienen que los caracteres moleculares son mejores:

- Taxonomía basada en ADN – Holotipos de ADN



Los mensajes son:

- conceptualmente un carácter molecular es lo mismo que un carácter morfológico
- cada clase de carácter tiene sus ventajas y desventajas
- el tema no pasa por la clase de carácter
- el tema pasa por como se eligen y se registra la variación de los caracteres

el problema esta en los sistemáticos...

Morfología vs. Moléculas

Caracteres morfológicos

Ventajas:

- estudio de especímenes depositados en colecciones
(previo a colecciones de tejidos)
- estudio de especímenes fósiles (ADN antiguo)
- uso de información ontogenética
- costo (?)



Caracteres morfológicos

Desventajas:

- taxones divergentes pueden tener pocos caracteres en común
- efecto ambiental difícil de disecar
- convergencia adaptativa (??)

¿reflejo de nuestro mayor conocimiento de morfología funcional?

- endotermos independientemente han enriquecido su genoma en GC
... quizás no sea adaptativo, sino sesgos sustitucionales

Caracteres moleculares

Ventajas:

- no hay efecto ambiental; estrictamente material hereditario
- descripción de los estados de carácter no es ambigua
(es "A" o "G" no "fino" y "menos fino")
- es relativamente fácil generar una matriz de miles de caracteres
- ciertos genes homólogos existen en todos los taxa
- regiones diferentes del genoma evolucionan con distinta tasa
permitiendo estudiar problemas filogenéticos a distintos niveles

Caracteres moleculares: ventajas...



Caracteres moleculares

Ventajas:

- En muchos casos podemos asumir que la tasa de cambio es constante en el tiempo

J. Nature Biol (1968) 4, 371-380

Surge la idea del 'reloj molecular'

Molecular as Documents of Evolutionary History
 Peter Doolittle and Lynn Margulis

Ejemplo con la α -globina

Figure 8.7. Relation between estimated number of amino acid substitutions in cytochrome b_5 between species in Figure 8.6, against time since each pair diverged from a common ancestor. The straight line is expected based on a uniform rate of amino acid substitution during the entire period. (From Kimura 1968)

Caracteres moleculares: Concepto de reloj molecular

- derivado inicialmente de observaciones como la del ejemplo de la α -globina

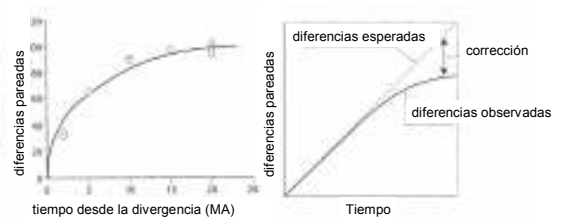
- relación lineal entre tiempo de divergencia entre especies y "distancia molecular" entre ellas

Esta distancia puede ser:

- la observada
- o estimada (corrección basada en un modelo)

Esta distancia puede ser:

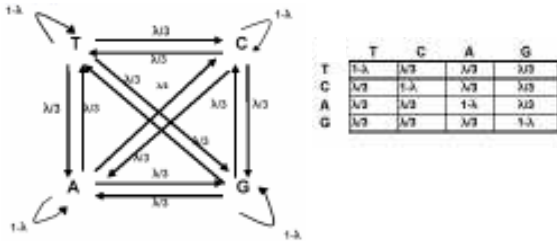
- la observada
- o estimada (corrección basada en un modelo)



Las diferencias observadas no corresponden muchas veces con las esperadas debido a los "cambios sobre cambios"

Distancia estimada: Modelo de Jukes & Cantor (1969)

Este modelo asume que todas las sustituciones, para cualquier sitio, tienen la misma probabilidad de suceder



	T	C	A	G
T	1-λ	λ/3	λ/3	λ/3
C	λ/3	1-λ	λ/3	λ/3
A	λ/3	λ/3	1-λ	λ/3
G	λ/3	λ/3	λ/3	1-λ

q = proporción de nucleótidos idénticos entre dos secuencias

$$q_{t+1} = (1 - 2\lambda)q_t + \frac{2}{3}\lambda(1 - q_t)$$

Probabilidad de que no haya sustituciones a $t+1$

Probabilidad de que a tiempo a t dos nucleótidos sean diferentes para el mismo sitio, y se tornen idénticos en $t+1$

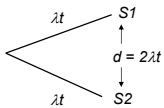
Usando un modelo de tiempo continuo podemos describir la ecuación anterior como:

$$\frac{dq}{dt} = \frac{2\lambda}{3} - \frac{8\lambda}{3}q$$

La solución a esta ecuación con condiciones iniciales $q = 1$ y $t = 0$

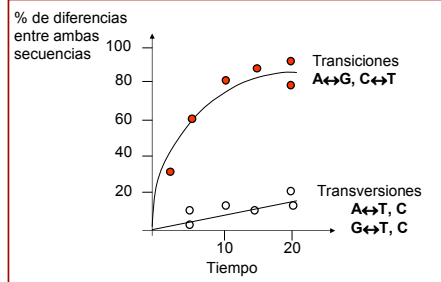
$$q = 1 - \frac{3}{4}(1 - e^{-8\lambda t/3})$$

Bajo este modelo, el número de sustituciones esperadas por sitio entre dos secuencias (d) es: $2\lambda t$, siendo t el tiempo de divergencia entre ambas secuencias



$$d = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right)$$

...en donde $p = 4q$

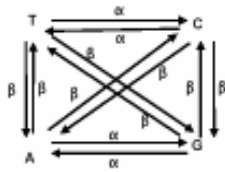


Al comparar dos secuencias, vemos que normalmente el número de transiciones es mucho más alto que el de transversiones

Para genes nucleares el ratio transiciones/transversiones oscila entre 0,5- 2, pero en genes mitocondriales puede ser tan alto como 15.

(Vigilant, 1991; en Nei & Kumar, 2000)

Distancia estimada: Modelo de 2 parámetros de Kimura



Distinta probabilidad si la sustitución es una transición (α : A↔G, C↔T) o una transversión (β : A↔T, C; G↔T, C)

	T	C	A	G
T	$1-\alpha-2\beta$	α	β	β
C	α	$1-\alpha-2\beta$	β	β
A	β	β	$1-\alpha-2\beta$	α
G	β	β	α	$1-\alpha-2\beta$

Definimos:

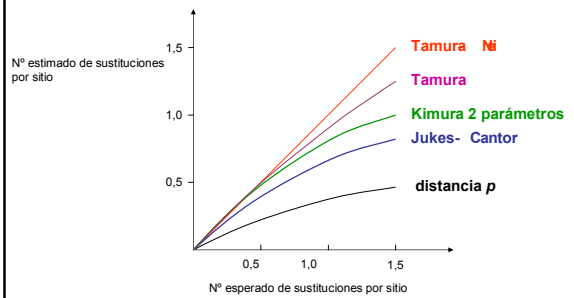
O = pares idénticos

P = pares transicionales

Q = pares transversionales

$$d = -\left(\frac{1}{2}\right)\ln(1-2P-Q) - \left(\frac{1}{4}\right)\ln(1-2Q)$$

Comparación de diferentes medidas de distancias para estimar el número de sustituciones:



Asumiendo un modelo de Tamura-Nei y $n = \infty$, modificado de Nei & Kumar (2000)

Recientemente se han desarrollado métodos más sofisticados, los cuales estiman la divergencia mediante otros parámetros:

Ejemplos:

por **Máxima Verosimilitud** (e.g. Rambaut & Bromham, 1998):

se toma un modelo de la evolución de la secuencia (un grupo de parámetros que describen el patrón de sustitución) y se busca la combinación de valores de esos parámetros que den el valor más alto de probabilidad de obtener esa secuencia

por **inferencia Bayesiana** (e.g. Kishino, *et al.* 2001):

éste método selecciona el agrupamiento (árbol) que tenga la probabilidad más alta de ser correcto, bajo un modelo específico de sustitución nucleotídica

Estos métodos de corrección de distancia comparten varias asunciones:

- Todos los sitios cambian en forma independiente
 - La tasa de sustitución es constante a lo largo del tiempo y entre los diferentes linajes
 - Todas las secuencias tienen la misma frecuencia de bases (equilibrio de composición)
 - Las probabilidades condicionales de sustitución nucleotídica son las mismas para todos los sitios y no varían a lo largo del tiempo
- Todas estas asunciones hacen a los métodos operativos, pero en muchos casos no se ajustan a la realidad...

Caracteres moleculares

Desventajas:

Comunes a cualquier tipo de carácter:

- Dificultad en el establecimiento de homólogas
- Niveles de variación inadecuados para el problema a tratar

lo que lleva a muchos autores a **asumir hipótesis *ad hoc***
(e.g., pesar caracteres, incorporar modelos de evolución)

Caracteres moleculares

Desventajas (más):

Propios de los caracteres moleculares:

Árbol de gen / árbol de las especies

fijación diferencial de polimorfismos
paralogía en familias multigénicas
transferencia horizontal
genes duplicados en poliploides
seudogenes

recombinación en genes nucleares

tiempo y \$\$\$\$

Ejemplo de árbol de gen distinto del árbol de las especies

Topología mejor corroborada

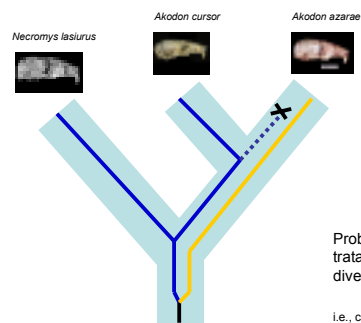


Topología reconstruida



¿Qué es lo que puede estar pasando?

Fijación alternativa de polimorfismos ancestrales



Problema mayor cuando se
tratan de reconstruir
divergencias "recientes"

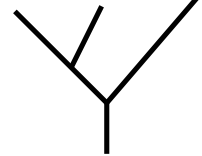
i.e., cuando los internodos son cortos

Otro ejemplo; otra causa

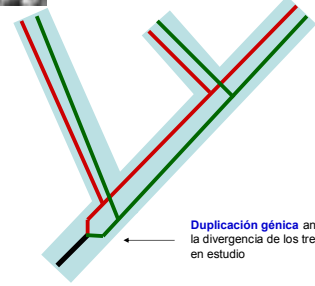
Topología mejor corroborada



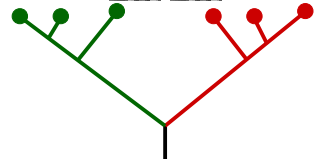
Topología reconstruida



Árbol de las especies
Árbol de los genes



Duplicación génica anterior a la divergencia de los tres taxa en estudio

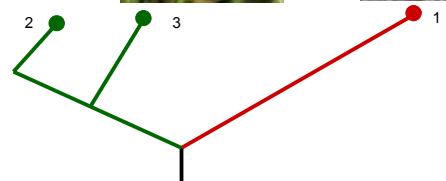


Las copias 1, 2, y 3 del gen ● son homólogos ortólogos

(idem respecto a las 3 copias de ●)

● y ● son homólogos parálogos; hay dos copias del gen ● en el genoma
(no estamos hablando de los dos alelos del gen ● que existen en un genoma diploide, estamos hablando de que el gen ● se ha duplicado)

Pensar en las familias multigénicas



Estos fueron los alelos estudiados
(o 2 y 3 ● y 1 ●)

MENSAJE: estudio filogenético debe de basarse en genes homólogos ortólogos

Consideraciones finales

- La sistemática molecular es cada vez más popular
recordar: son pocos los estudios que son 100% moleculares
- Los caracteres moleculares:
 - no tienen nada de especial
 - no son intrínsecamente mejores ni peores que los caracteres morfológicos
 - respecto a estos tienen ciertas ventajas y desventajas
 - algunos problemas son comunes a cualquier clase de carácter
 - otros son propios

Considerando que:



- la topología con mayor apoyo será aquella que pasa la mayor cantidad de pruebas (tests), y
- que cada carácter constituye una prueba de hipótesis (la topología) independiente:

Lo mejor es combinar en un mismo análisis caracteres de distintas fuentes