BIOLOGÍA MOLECULAR

PRUEBA DE EVALUACIÓN CONTINUA 2

MÁSTER BIOINFORMÁTICA Y BIOESTADÍSTICA

Santiago Royuela Samit

Enero 2022

Ejercicio 2. Imaginemos que la siguiente secuencia:

0. ATGGCACGCACAATAAACTGCGCGAACACAGCGCGCGCGCCCTGAGCTATA

Contiene la región codificadora completa del alelo normal de un gen hipotético que no contiene ningún intrón. Este fragmento de ADN comienza con el **codón ATG** de **inicio de la traducción**, contiene **el gen completo** y se extiende algunos nucleótidos (nt) más por la región UTR hasta completar un **total de 51 nt**. Al obtener la secuencia de este gen en varios individuos de una población natural, se han encontrado algunos individuos que mostraban variabilidad en esta secuencia. En concreto, se han encontrado las siguientes variantes que siempre hemos cortado los 51 nt de su inicio.

1. ATGGCAAGCACAATAAACAGCGCCAACTCAGCGCGGGCGCCCTGAGCTATA

2. ATGGCACGCACAATAAACTGCGCGAACACAGCGCGCGCGCCGTGAACTTTA

3. ATGGCACGGACAATAAACAGCGCGAACACAGCGCGCGCGCCCTGATCTATT

4. ATGGCCCGGACAATAAACTGAGCCAACACAGCGCGGGCGCCCTGAGCTATA

5. ATGGCACGGACAATCAATTGCGCGAATACAGCGCGGGCGCCGTGAGCTATA

6. ATGGCGAGCACAATCAACTGCGCGAACACACGCGCGCGCCCTGAGCTATAA

7. ATGGCACGGACAATAAACAGTGCGAACACAGCGCGCGCGCCCTGATCTATT

8. ATGGCAAGCACAATAAACTGCGCCAATTCAGCGCGGGCGCCCTGAGATATA

a) Ordena estas secuencias (de menor a mayor efecto) según pienses que será su grado de efecto sobre el fenotipo de los individuos portadores de estas secuencias.

b) Explica el porqué de esta ordenación.

Suponemos que nos han dado la cadena codificadora en su sentido 5'->3'. Primero analizamos la **cadena 0** del alelo normal del gen hipotético.

Si hacemos la transcripción a la secuencia de ARN obtenemos la cadena:

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGA**GCUAUA

Primero tenemos el codón de inicio (AUG), que codifica para Metionina. El codón de **STOP** se ubica las posiciones de la cadena: 42,43 y 44 (UGA). El resto de nucleótidos hasta los 51 constituyen la región 3' UTR (GCUAUA). Hay 42 codones codificantes para aminoácidos.

Y la proteína que se obtiene tras la traducción es: **MARTINCANTARAP**(STOP). Tras el codón de STOP hay 2 codones más no codificantes y que pertenecen a la región 3' UTR. Miramos las posiciones sinónimas y no sinónimas de la cadena (sin contar los codones de inicio y STOP, ni la región 3'UTR. Lo paso por mi programa de python):

Número de posiciones sinónimas:

10.833333333333332

Número de posiciones No sinónimas:

28.166666666666668

Pasamos a analizar el resto de secuencias conforme al orden que nos piden:

• Cadena 2: ATGGCACGCACAATAAACTGCGCGAACACAGCGCGCGCGCCGTGAACTTTA

Transcripción:

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCG**UGAACUUUA**

Traducción: **MARTINCANTARAP**(STOP). Vemos que la proteína a codificar no se ve alterada en su estructura primaria o secuencia de aminoácidos, así que esta mutación o secuencia poco afectará al fenotipo al codificar para la misma secuencia de aminoácidos. La estructura secundaria y terciaria tampoco se verán afectadas, puesto que la secuencia de aa es la misma.

Analizamos los cambios entre estas dos cadenas, la 0 y la 2:

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGAGCUAUA**

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCC**G**UGA**A**CU**U**UA

Cambios entre nucleótidos: hay 3 nucleótidos distintos. 1 cambio es una transición y 2 son transversiones. Posiciones: [42, 46, 49]. Sólo una mutación tiene lugar en la región codificadora, pero da lugar a un codón sinónimo que traduce para el mismo aminoácido. Las otras 2 mutaciones se dan en la región 3'UTR que no afecta a la traducción.

• Cadena 5: ATGGCACGGACAATCAATTGCGCGAATACAGCGCGGGCGCCGTGAGCTATA

Transcripción:

AUGGCACGGACAAUCAAUUGCGCGAAUACAGCGCGGGCGCCGUGAGCUAUA

Traducción: **MARTINCANTARAP**(STOP). Vemos que la proteína a codificar no se ve alterada en su estructura primaria o secuencia de aminoácidos, así que esta mutación o secuencia poco afectará al fenotipo al codificar para la misma secuencia de aminoácidos, manteniéndose las estructuras secundaria y terciaria.

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGAGCUAUA**

AUGGCACG**G**ACAAU**C**AA**U**UGCGCGAA**U**ACAGCGCG**G**GCGCC**G**UGAGCUAUA

Cambios entre nucleótidos: hay 6 nucleótidos diferentes, 4 transversiones y 2 transiciones. Posiciones: [9, 15, 18, 27, 36, 42]. En este caso, todas las mutaciones tienen

lugar en la región codificadora, dando lugar a codones sinónimos. Además, las mutaciones se dan en la tercera posición de los codones.

• Cadena 3: ATGGCACGGACAATAAACAGCGCGAACACAGCGCGCGCGCCCTGATCTATT

Transcripción:
AUGGCACGGACAAUAAACAGCGCGAACACAGCGCGCGCGCCC**UGA**UCUAUU

Traducción: **MARTINSANTARAP**(STOP). Vemos que hay un cambio en solo un aminoácido (el codificado por el 7º codón), de C a S en la séptima posición de la estructura primaria de la proteína, ello conllevará a cambios en las estructuras secundaria y terciaria, por lo que afectará al fenotipo con mayor grado que las anteriores vistas.

Analizamos los cambios entre estas dos cadenas, la 0 y la 3:

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGA**GCUAUA

AUGGCACG**G**ACAAUAAAC**A**GCGCGAACACAGCGCGCGCGCCC**UGA**U**C**UAU**U**

Cambios entre nucleótidos: hay 4 nucleótidos distintos, todo trasnversiones. Posiciones: [9, 19, 46, 51]. Hay 2 mutaciones en la región codificadora, una de ellas no sinónima que provoca el cambio de aminoácido en la secuencia de la proteína.

• Cadena 7: ATGGCACGGACAATAAACAGTGCGAACACAGCGCGCGCGCCCTGATCTATT

Transcripción:
AUGGCACGGACAAUAAACAGUGCGAACACAGCGCGCGCGCCCUGAUCUAUU

Traducción: **MARTINSANTARAP**(STOP). El número de aminoácidos no ha cambiado, pero sí uno de ellos, el 7º aa. La estructura primaria de la proteína se ve modificada por 1 aminoácido.

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGA**GCUAUA
AUGGCACG**G**ACAAUAAAC**A**G**U**GCGAACACAGCGCGCGCGCCC**UGA**U**C**UAU**U**

Cambios entre nucleótidos: hay 5 nucleótidos distintos, todo transversiones. Posiciones: [9, 19, 21, 46, 51]. Hay 3 mutaciones en la región codificadora y 2 en la 3'UTR. De las 3 en la codificadora, 2 son sinónimas y otra no lo es, dando lugar a una estructura primaria de la proteína distinta en un aminoácido, cosa que afectará a sus estructuras secundaria y terciaria.

• Cadena 1: ATGGCAAGCACAATAAACAGCGCCAACTCAGCGCGGGCGCCCTGAGCTATA

Transcripción: **UG**GCAAGCACAAUAAACAGCGCCAACUCAGCGCGGGCGCCC**UGA**GCUAUA

42 codones codificantes, siendo el primero el de inicio y codificando para Metionina. Un codón de STOP y 6 nucleótidos en la región 3'UTR

Traducción: **MASTINSANSARAP.**(STOP). Vemos que la proteína a codificar se ha visto modificada en 3 aminoácidos (el 3º,7º y 10º), lo que supondrá un cambio funcional y una repercusión en el fenotipo. El número de codones traducidos es el mismo que en la cadena 0, de 42 codones codificantes para aminoácidos y uno de STOP. El resto hasta 51 s la región 3'UTR.

Analizamos los cambios entre estas dos cadenas, la 0 y la 1:

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGAGCUAUA**

AUGGCA**A**GCACAAUAAAC**A**GCGC**C**AAC**U**CAGCGCG**G**GCGCCC**UGAGCUAUA**

He diseñado un código en Python (código 1) que se adjunta en el Anexo para detectar las diferencias entre ambas cadenas, en qué posiciones y sin son transversiones o transiciones. Los resultados son:

Cambios entre nucleótidos: hay 5 diferencias entre los nucleótidos de las cadenas. Todos los cambios son transversiones, siendo las posiciones [7, 19, 24, 28, 36]. Recordemos que las transversiones son menos probables que las transiciones.

Las 5 mutaciones se dan en la región codificadora, 3 no son sinónimas, dando lugar a la traducción de otros aminoácidos, y las otras 2 sí son sinónimas. A pesar de haber 3 aminoácidos diferentes en la estructura primaria de la proteína, el número de aminoácidos de la cadena proteica se mantiene, pero se verán afectadas en mayor grado que en las anteriores las estructuras secundaria y terciaria.

• **Cadena 4:** ATGGCCCGGACAATAAACTGAGCCAACACAGCGCGGGCGCCCTGAGCTATA

Transcripción:

AUGGCCCGGACAAUAAAC**UGAGCCAACACAGCGCGGGCGCCCUGAGCUAUA**

Traducción: **MARTIN**(STOP). La traducción se detiene antes de terminar la secuencia de aa de la estructura primaria de la proteína. El fenotipo se verá altamente afectado, pues faltan 8 aminoácidos para completar la estructura primaria de la proteína.

Analizamos las diferencias entre estas dos cadenas, la 0 y la 4:

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGAGCUAUA**

AUGGC**C**CG**G**ACAAUAAACUG**A**GC**C**AACACAGCGCG**G**GCGCCCUGAGCUAUA

Cambios entre nucleótidos: hay 5 nucleótidos diferentes, todos son transversiones. Posiciones: [6, 9, 21, 24, 36]. Las 5 mutaciones tienen lugar en la región codificadora y se ha dado el caso que todas son sinónimas menos una, que es un STOP y detiene la traducción. Si este codón STOP que se ha formado volviese a mutar para codificar a C, la secuencia recuperaría la estructura primaria de la proteína, pero mientras tanto, ésta ha perdido gran parte de su secuencia de aminoácidos, aunque conservando el orden en los que permanecen. Por tanto, la estructura secundaria y terciaria se verán afectadas.

• **Cadena 6:** ATGGCGAGCACAATCAACTGCGCGAACACACGCGCGCGCCCTGAGCTATAA

Transcripción:

AUGGCGAGCACAAUCAACUGCGCGAACACACGCGCGCGCCCUGAGCUA**UAA**

Traducción: **MASTINCANTRARPEL**(STOP). El cambio en la proteína sintetizada es debido a la secuencia y número de aminoácidos. La del gen normal codifica para 14 aminoácidos, y este gen codifica para 16 aminoácidos, con lo que varía mucho su estructura primaria, tanto en secuencia o tipos de aminoácidos, así como en su número. Por tanto, los aspectos funcionales de la proteína, o su fenotipo, se verá altamente afectado.

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGA**GCUAUA
AUGGC**GA**GCACAAU**C**AACUGCGCGAACACA**CGCGCGCGC**CCU**GAGCUAUAA**

Cambios entre nucleótidos: hay 21 cambios de nucleótidos entre las dos secuencias en las posiciones: [6, 42, 44, 45, 47] para transiciones y [7, 15, 31, 32, 33, 34, 35, 36, 37, 38, 39, 43, 46, 48, 49, 50] para transversiones. Al contrario que la cadena 5, ésta es más difícil que vuelva a recuperar su estructura primaria.

c) ¿Cuántas secuencias nucleotídicas diferentes podrían codificar la misma proteína que la secuencia nº 0 del enunciado?

Si hacemos la transcripción a la secuencia de ARN obtenemos la cadena:

AUGGCACGCACAAUAAACUGCGCGAACACAGCGCGCGCGCCC**UGA**GCUAUA

Primero nos fijamos en que tenemos 6 posiciones nucleotídicas en la región 3'UTR después del codón de STOP. Por tanto, si estos nucleotídicos cambiaran, no afectaría a la traducción de aminoácidos de la proteína, luego tenemos, por un lado:

4x4x4x4x4x4 = 4.096 cadena distintas, por ahora.

El primer codón de inicio/Metionina no puede verse alterado, luego tenemos 3 posiciones, o el primer codón, que no pueden variar. De otro lado, la traducción debe terminar con un codón de STOP, por lo que habrá 3 posibles, uno por cada codón de STOP de los 3 que hay. Ahora queda ver de cuántas maneras, mediante codones, podemos traducir ARTINCANTARAP.

A → 4 CODONES ; R → 6 CODONES ; T → 4 CODONES ; I → 3 CODONES

N → 2 CODONES ; C → 2 CODONES ; A → 4 CODONES ; N → 2 CODONES

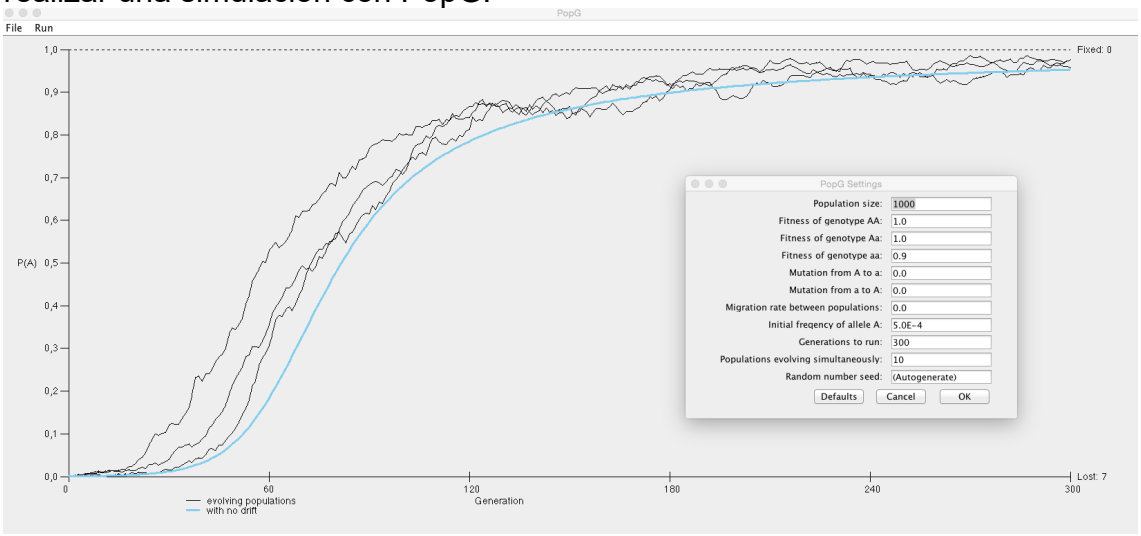
T → 4 CODONES ; A → 4 CODONES ; R → 6 CODONES ; A → 4 CODONES

P → 4 CODONES

En total, el cálculo sale: 4.096 x 3 x 4^7 X 6^2 X 3^1 x 2^3 = 173,946,175,488 (las "," son valores posicionales y no punto decimal), que es del orden de 174 mil millones de cadenas nucleotídicas.

Ejercicio 3.

La siguiente figura muestra los parámetros introducidos y el output obtenido al realizar una simulación con PopG.



a) comenta los parámetros de entrada que se han utilizado. En la generación 0, ¿cuantos individuos de cada genotipo (AA, Aa, aa) hay? ¿Qué escenario se está simulando? ¿Qué fuerzas evolutivas están actuando sobre las poblaciones simuladas?

b) Comenta el resultado obtenido. ¿Por qué no se comportan todas las poblaciones del mismo modo?

Responderé comentando la simulación que se nos presenta:

Con estos parámetros estamos haciendo una simulación simultánea de 10 poblaciones (en paralelo e independientes, digamos) de 1.000 individuos bajo los mismos parámetros que ahora comentamos y que se aparearán durante 300 generaciones. Estamos haciendo 10 experimentos, digamos, pues la evolución será siempre "azarosa", al margen de si hay o no una selección, sobre la que fluctuará este azar. La deriva en principio, al tratarse una población grande, no debería ser una fuerza evolutiva, presuponiendo, en principio, un equilibrio del tipo Hardy-Weinberg en el instante 0. Este equilibrio, como veremos, se verá modificado a causa de las eficacias biológicas que se indican en la simulación; por tanto, no tendremos un equilibrio de Hardy-Weinberg y las frecuencias alélicas evolucionarán hasta dos puntos de equilibrio posibles, pues la coexistencia de alelos, como veremos, no es viable a largo plazo –no es de equilibrio; la extinción de no de los alelos será condición necesaria para el equilibrio de esta población-.

Primero observamos que hay un genotipo concreto con una eficacia biológica inferior al resto, lo cual irá en su detrimento. El genotipo es el homocigótico aa, que tiene una eficacia de 0.9 relativa a 1 con respecto al resto de genotipos. Podemos decir

que la selección natural irá en detrimento de este genotipo, el cual solo podrá mantener sus frecuencias o aumentarlas a causa de la deriva o azar, pero no a causa de una selección que la encamine al éxito. Por el contrario, los otros genotipos evolucionarán con una eficacia mayor a la anterior, favoreciendo el aumento de la frecuencia del alelo A en los genotipos de la población, pero jugando con el hándicap de una frecuencia inicial baja del alelo "A" y que la deriva `podrá subsumir.

En el modelo no hay mutaciones ni migraciones, lo cual nos permite, salvo la eficacia que trataremos más adelante, suponer en primera aproximación un equilibrio de Hardy-Weinberg en el punto inicial, el cual se verá roto a causa de las eficacias biológicas en donde el alelo A está presente.

Ahora realizamos un cuadro teniendo en cuenta las eficacias biológicas y/o los coeficientes de selección. Según los parámetros, tenemos una selección en contra del homocigoto aa:

Genotipo	**AA**	**Aa**	**aa**	**Total**
Frecuencia Inicial Genotípica	AA (p^2) = 25 X 10^{-8}	Aa(2pq) = 0,0009995	aa(q^2) = 0,99900025	1
w (eficacia)	1	1	1-s = 0.9	
S (coeficiente de selección)	0	0	s = 0.1	
Frecuencia tras la selección	P^2=25 X 10^{-8}	2pq = 0,0009995	q^2(1-0.1)= $0.9q^2$ = 0.899100225	1-($0.1q^2$) = 0.900099975

En el momento inicial t = 0 tenemos una población de 1.000 indivíduos, lo que supone 2.000 alelos. Si multiplicamos por la frecuencia inicial del alelo "A" ($5x10^{-4}$) nos sale 1 sólo alelo A. Si hacemos lo propio con "q" y el alelo "a", obtenemos 1999 alelos "a". Como hemos son mil individuos, tenemos que 999 tienen genotipo "aa", y solo 1 tiene genotipo "AA".

Ahora veamos cómo evolucionan en el tiempo las frecuencias génicas (la de los alelos "a" y "A"):

$$q' = \frac{(1-s)q^2 + pq}{p^2 + (1-s)q^2 + 2pq} = \frac{0.9q^2 + pq}{1 - 0.9q^2}$$

$$p' = \frac{p^2 + pq}{p^2 + (1-s)q^2 + 2pq} = \frac{p^2 + pq}{1 - 0.9q^2}$$

A primera vista en las ecuaciones, ambas están divididas por el mismo factor. En el numerador ambas tienen el factor pq sumando y, mientras p' crece en proporción a p^2, q' lo hace en proporción a q^2 multiplicado por 0.9, que es menor que 1.

Tenemos que:

$$\Delta p = p' - p = \frac{0.1pq^2}{1 - 0.1q^2}$$

El equilibrio, en donde $\Delta p = 0$, solo se alcanzará si, o bien p, o bien q, son iguales a 0. No puede existir un equilibrio en donde coexistan los dos alelos. En nuestro caso, como solo hay un coeficiente de selección s, contra el homocigoto "aa", el equilibrio sólo se alcanza para p=0.1/0.1=1 y q=0, o bien p se hace cero debido a que, por la deriva, no prolifera y se extingue. De hecho, en t=0 tenemos un solo individuo heterocigótico "Aa" que, al aparearse con un individuo "aa" podría dar lugar a una descendencia de sólo individuos "aa". Desconozco, en el modelo del programa, cuánta descendencia se tiene en cada generación, y si hay o no mortandad, así como el ritmo de apareamiento.

En nuestro caso las fuerzas que actúan sobre las poblaciones simuladas son la selección en contra del homocigoto "aa", que favorecerá los genotipos "AA" y "Aa", en definitiva, al alelo A. De otro lado, tendremos un efecto debido al azar puro, la deriva, que no tiene una dirección privilegiada y que decrecerá según $\frac{1}{\sqrt{N}} \sim 0.0578$ en nuestro caso.

En el ejemplo que se nos presenta, tenemos que hay 3 poblaciones de 10 simuladas, en las que la frecuencia génica del alelo A irá camino de fijarse a 1 tras las 300 rondas/apareamientos simuladas. Si dejásemos más generaciones, estas 3 que se muestran en la figura acabarían con una probabilidad casi del 100% fijando el alelo "A" con la extinción total del "a" (solo una catástrofe o conjuración de probabilidades a causa de la deriva evitaría este proceso de fijación final). Las otras 7 simulaciones, en el efecto de la deriva y por abrirse camino en con ella en la población desde una frecuencia inicial muy baja, sólo un heterocigoto de mil, éste alelo "A" se ha extinguido sin poder prosperar.

Las diferencias en las 3 gráficas son a causa del mero azar en la simulación, que a la vez dan cuenta o emulan la deriva génica. Ambas gráficas empiezan con un crecimiento muy lento que, si lo superan, pasan a una fase de inflación o crecimiento exponencial, hasta acercase a una frecuencia cercana a 0.5, donde refleja un comportamiento con crecimiento logarítmico hasta saturarse en 1. Si no hubiese deriva alguna, una población infinita, la línea azul sería el destino de la frecuencia génica del alelo "A". Las curvas mostradas en el gráfico, fluctúan entorno a esta curva azul ideal, a causa de los efectos del azar o deriva.

Si uno toma esos parámetros y simula para una población mayor, de 10.000, observa que hay más poblaciones de las 10 simuladas que tienden a fijar el alelo A y que sobreviven sin extinguirse, pues los efectos de deriva se van "diluyendo" y, si fuera infinita, siempre se fijaría el alelo A tras las pertinentes generaciones en todas las poblaciones simuladas.

Ejercicio 4. Un investigador estudia la divergencia entre dos especies muy cercanas y para ello analiza la secuencia de un fragmento del segundo exón codificador de un gen que tiene cuatro exones codificadores. Dicho fragmento cuenta con **183 codones** y al

considerar las secuencias de las dos especies ha contabilizado **400,66 posiciones no sinónimas**. Por otro lado, las únicas diferencias que ha observado entre los codones de estas dos secuencias son los que aparecen en la tabla siguiente:

a) ¿Cómo puede encontrar un número no entero de posiciones no sinónimas?

Tenemos dos exones concretos de dos especies cercanas en divergencia. Este exón que analizamos es el segundo codificador del gen, por tanto, no hay codón de inicio ni de STOP, así que los 183 codones serán codificantes. En total, el exón contiene 549 pares de nucleótidos.

Para calcular el número de posiciones no sinónimas se divide la secuencia de ADN en codones o tripletes. En cada codón podemos originar 3 cambios, para cada uno de sus nucleótidos. Una vez efectuado el cambio en la posición "i" (i=1..3) del codón, éste pasa a variar su codificación, siendo que puede dar lugar a otro codón que codifica para el mismo aminoácido (cambio sinónimo), que codifique para otro aminoácido (cambio no sinónimo), o que codifique para STOP (no se tiene en cuenta). Si hay un cambio sinónimo, sumamos un 1 a las posiciones sinónimas del nucleótido, si hay cambio no sinónimo, lo sumamos a las posiciones no sinónimas y, si es un cambio hacia STOP, entonces no hacemos nada. Pero, como son 3 posiciones por codón a considerar, deberemos dividir estos números entre 3 cuando los sumemos.

Bien, cuando un nucleótido concreto, el de la posición "i", de un codón cambia, puede hacerlo hacia 3 bases posibles, las distintas a la suya. Ello originará 3 opciones: que el codón con ese nucleótido cambiado codifique para el mismo aminoácido (sinónima), que codifique para otro aminoácido (no sinónima) o que codifique para STOP (no se tendrá en cuenta). Es decir, que por cada posición nucleotídica del codón tenemos 3 cambios que pueden dar lugar al mismo aminoácido, a otro distinto, o a un STOP y no ser considerado. Y así se irá sumando para cada posición nucleotídica del codón, siendo que, como algunos cambios son sinónimos y otros no sinónimos, e incluso algunos no se suman al ser STOP's, al dividir por 3 las posiciones que han sido no sinónimas, puede resultar un número no entero.

Insistiendo. Cada nucleótido de un codón puede cambiar a 3 nucleótidos diferentes. Como se ha de dividir por 3 en cada posición nucleotídica del codón, si todos los cambios fuesen no sinónimos, tendríamos hasta un máximo de 1 en esa posición no sinónima del codón, pero si uno de los cambios de bases produce el mismo aminoácido y los otros 2 no, entonces habrán 2/3 posiciones no sinónimas a sumar al total que se va acumulando conforme recorremos los codones del exón.

Una vez se han sumado las posiciones no sinónimas de todos los codones de una de las cadenas, se procede igualmente con la segunda para obtener la media, como veremos.

b) ¿Cuántas diferencias sinónimas y cuántas no sinónimas hay entre estas

secuencias? Explícalo.

Como las dos secuencias son iguales salvo en los codones que nos indican en la tabla, vamos a calcular el número de diferencias sinónimas y no sinónimas. Para los codones que son iguales, no hay diferencias ni sinónimas ni no sinónimas, así pues, analizamos las que son diferentes e indicadas en la tabla adjunta, que nos darán los totales de las diferencias sinónimas y no sinónimas. Para agilizar los cálculos, utilizo el programa diseñado en Python, que me dice las diferencias sinónimas y no sinónimas, así como los aminoácidos para los que codifica cada codón.

Nº Codón	8	9	19	30	51	59	60	64	65	75	79	98	99	111	134	147	155
Sec. A	CTA	TTA	TCT	GCG	CCA	GGT	TCA	CAA	CTA	AAT	TTT	GGA	CTA	AAA	GGT	CGA	TCG
Aminoácido	L	L	S	A	P	G	S	Q	L	N	F	G	L	K	G	R	S
Sec. B	CTG	TTG	TCC	GCC	CCG	GGC	TCG	CAG	GTA	AAC	TTC	GGG	TTG	AAG	GGG	AGA	TCC
Aminoácido	L	L	S	A	P	G	S	Q	V	N	F	G	L	K	G	R	S
S_d	1	1	1	1	1	1	1	1	0	1	1	1	2	1	1	1	1
n_d	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Tras el recuento, tenemos que hay 17 diferencias sinónimas y 1 no sinónima:

$$S_d = 17 \text{ y } N_d = 1$$

c) Calcula p_S y p_A. Calcula K_S y K_A. (Usar 5 decimales)

Si las cadenas tienen **183 codones** para el fragmento del segundo exón, en total contarán con 183*3 = **549 nucleótidos o pares de bases**. El número de posiciones sinónimas más el número de posiciones no sinónimas de una cadena debe ser igual al número de nucleótidos. Como el enunciado dice que al considerar las secuencias de las dos especies se han contabilizado **400,66 posiciones no sinónimas**, podemos saber el número de posiciones sinónimas. Recordemos que, al haber secuenciado los fragmentos de las dos especies, tenemos la media de las posiciones no sinónimas.

Salvo en estos codones que nos indican, los dos fragmentos de ADN son idénticos. Por tanto, en los codones idénticos las posiciones sinónimas y no sinónimas de la cadena serán las mismas. $S'_A = S'_B$ y $N'_A = N'_B$. Ahora calculamos las posiciones sinónimas y no sinónimas de la cadena "virtual" formada por los codones en los que ambas cadenas difieren:

Sec A: CTATTATCTGCGCCAGGTTCACAACTAAATTTTGGACTAAAAGGTCGATCG

Sec B: CTGTTGTCCGCCCCGGGCTCGCAGGTAAACTTCGGGTTGAAGGGGAGATCC

Cuando uno analiza el álgebra de las sinonimias en las posiciones de codones, se da cuenta que es asociativa y conmutativa, así que podemos analizar las posiciones sinónimas/no sinónimas de estas dos secuencias y sumarlas al total. Las paso por mi programa y obtengo:

Cadena A:

- 15,5 posiciones sinónimas.

- 35,5 posiciones no sinónimas.

Cadena B:

- 13,83333 posiciones sinónimas.
- 37.16666 posiciones no sinónimas

Como nos dan el total de posiciones no sinónimas, sumamos para estas:

$$\frac{(N'_A + 35{,}50000) + (N'_B + 37.16666)}{2} = 400{,}66000$$

Tenemos que $N'_A = N'_B$, despejando tendremos:

$N'_{A,B} = 364{,}32667$ posiciones no sinónimas en el fragmento que es igual en las dos cadenas y que no nos proporcionan. Por tanto, el número total de posiciones no sinónimas de los fragmentos totales será:

$N_A = 364{,}32667 + 35{,}50000 = 399{,}82667$ posiciones no sinónimas en la cadena A.

$N_B = 364{,}32667 + 37{,}16666 = 401{,}49333$ posiciones no sinónimas en la cadena B.

Cuya media es coincide con la dada por el enunciado de 400,66. Ahora, como sabemos el total de nucleótidos de los dos fragmentos, pasamos a calcular las posiciones sinónimas de ambos fragmentos totales:

$S_A = 549 - 399{,}82667 = 149{,}17333$

$S_B = 549 - 401{,}49333 = 147{,}50667$

Y podemos ver que $S_A + N_A = S_B + N_B = 549$, que es el número de nucleótidos o de pares de base, si se quiere. Ahora calculamos el número medio de posiciones sinónimas, pues el de no sinónimas ya nos lo dicen (aunque si lo hacemos sale el mismo resultado, claro está)

$S = \frac{149{,}17333 + 147{,}50667}{2} = 148{,}34$ posiciones sinónimas.

$N = \frac{399{,}82667 + 401{,}49333}{2} = 400{,}66$ posiciones no sinónimas.

Se podría hacer directamente, pero lo he hecho por pasos: S = 549 – 400,66 = 148,34 posiciones no sinónimas.

Calculamos las p's:

$$\boldsymbol{p_s = \frac{S_d}{S} = \frac{17}{148{,}34} = 0{,}114602}$$

$$\boldsymbol{p_A = \frac{N_d}{N} = \frac{1}{400{,}66} = 0{,}00250}$$

Ahora aplicamos la fórmula para las K's bajo el modelo de Jukes y cantor:

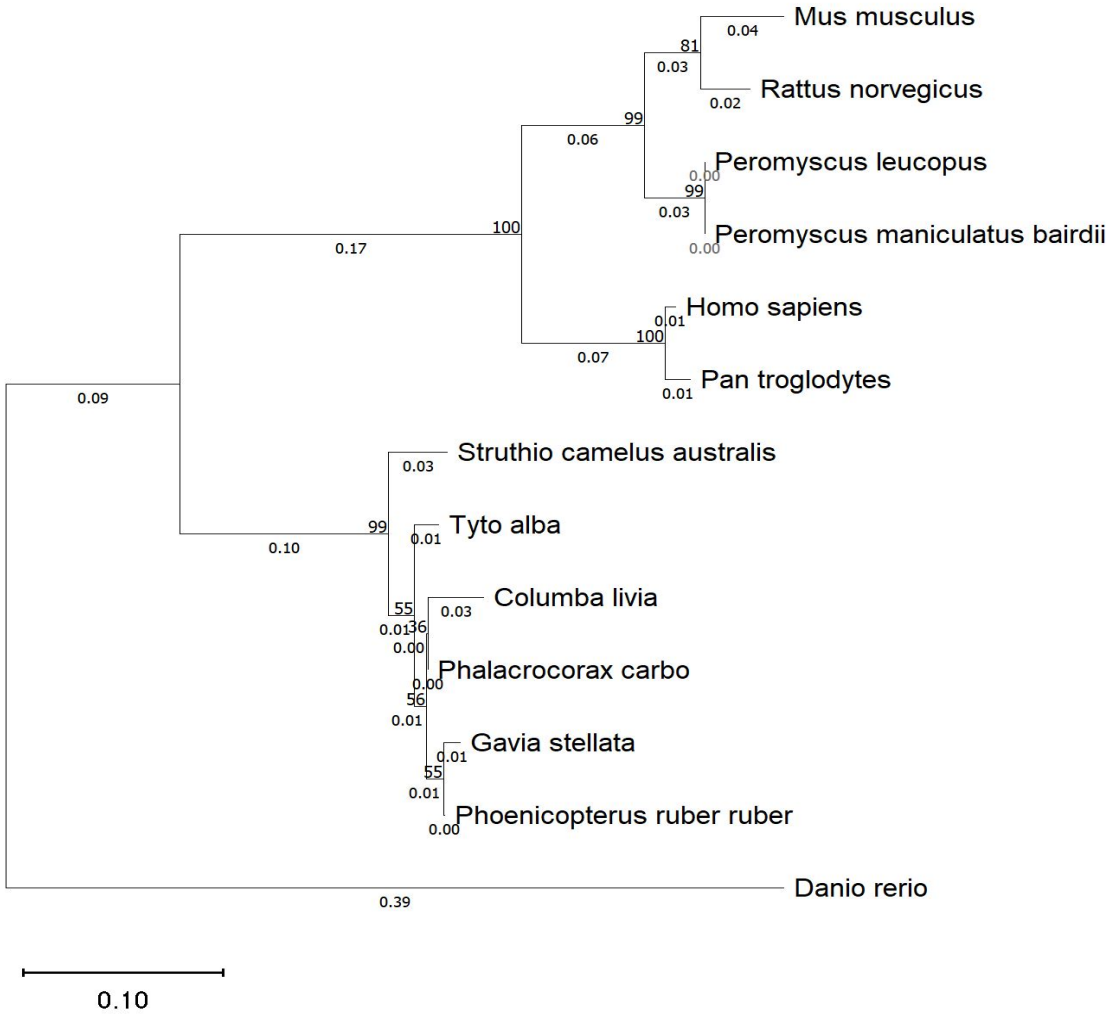
$$K_s = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_s\right) = 0,12437$$

$$K_A = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_A\right) = 0,00250$$

Ejercicio 5.

El archivo 13_especies.fas que encontrarás junto al enunciado de esta PAC contiene las secuencias aminoacídicas alineadas de una proteína x en 13 especies de animales.

a) Utilizando estas secuencias, construye un árbol Neighbour--Joining obteniendo los valores de bootstrap para 500 réplicas y pega la imagen obtenida a tu documento de respuesta a la PAC.



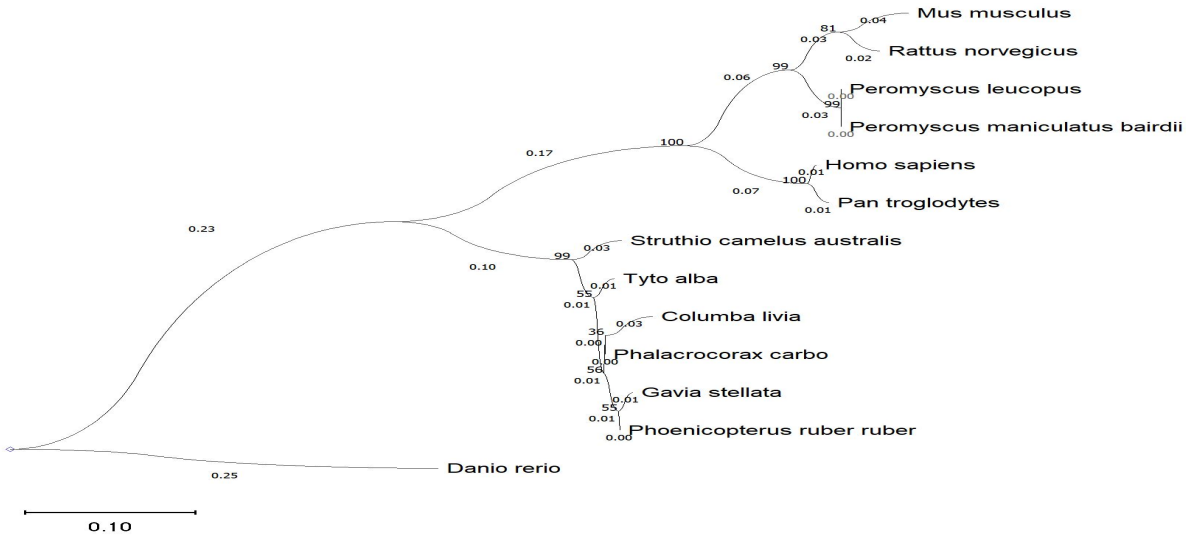
b) Viendo la imagen que has obtenido, ¿tienes la impresión que muestra unas relaciones filogenéticas correctas entre las especies representadas en el árbol? Razona tu respuesta. No entiendo de filogénetica más que lo estudiado en esta asignatura, así que diré lo que observo. Las 6 primeras especies desde arriba, de Mus muluscus a Pan troglodytes parecen estar bien ordenadas y soportadas con un buen bootsrap, lo que indica que las relaciones filogenéticas pueden ser correctas. Mirando

en wikipeia coinciden sus Reino, Filo, Clase y Familia o subfamilia, pero he de decir que yo nada entiendo de clasificaciones.

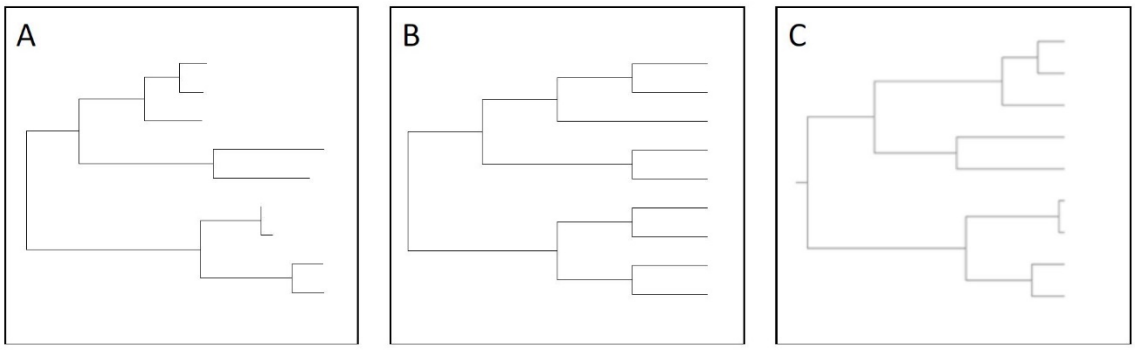
El OTH que se desprende de este grupo de OTU's y conecta con el nodo interno u OTH que va a parar al otro grupo de 6 especies u OTU's, parece bien soportado con un bootsrap alto, tanto el uno como el otro, lo que indica que ambos OTH proceden de un OTH anterior mediante ramas bien soportadas (100 y 99%). Hasta aquí, la clasificación sería correcta, en el sentido de las 6 primeras especies conectadas al OTH que dio lugar a las otras 6 especies, excepto la Danio rerio.

Dentro de este segundo grupo de 6 especies, vemos que las ramas ya no están bien soportadas, con un bootsrap muy mediano, lo que indica que puede que las relaciones no sean tan buenas como en las 6 anteriores. De otro lado, Danio rerio no tiene soporte de bootsrap, existiendo un UTH que le conecta basalmente a los dos grupos anteriores. En este sentido, cobra fuerza que sea una especie alejada de las otras agrupadas en dos grupos principales y, de esta manera, poder suponer que hubo un ancestro(UTH) común a Danio rerio y el resto de especies, enraizando el árbol de esta manera.

c) De las especies analizadas, ¿Qué especie es la que primero se separó de la evolución del resto de especies? El conocer este dato, ¿que nos permite hacer? La primera en separarse fue Danio rerio, en donde se indica en el árbol una distancia de su rama externa de 0.39, la mayor de todas, siendo ésta basal. Ello nos permite pasar de un árbol sin raíz a otro enraizado, introduciéndolo como un outgroup u OTU de referencia. Podemos considerar este OTU relacionado solamente de forma distante con las secuencias en estudio. De esta manera, podemos introducir la raíz en algún punto de la rama que une Danio rerio, como outgroup, al resto del grupo. Si lo hacemos en mega, podemos rotar en torno a la rama de Danio rerio y obtendremos un árbol enraizado en donde se han modificado la distancia de 2 ramas, la de Danio y la de un OTH que se separaron de un ancestro común, la raíz.



Ejercicio 5.



a) Indica qué tipo de árbol (cladograma, UPGMA Neighbour Joining) es cada uno de los tres gráficos y explica cómo has llegado a esa conclusión.

B es un cladograma, puesto que todas las especies actuales (OTU's) y sus ancestros (nodos internos) están alineados y las ramas no proporcionan información alguna más que su topología. C es un árbol tipo UPGMA, puesto que las especies actuales (OUT's) se alinean sobre el presente, y en donde cada una de las ramas tendrá una longitud en base a una evolución temporal bajo la hipótesis de igual ritmo de evolución o tasa de sustitución en las secuencias. En B todos los nodos internos están alineados, pero en C no, lo que indica, en éste último, distancias "temporales" y árbol ultramétrico. Por descarte y apariencia, A es un árbol tipo Neighbour Joining, pues las especies actuales no se sitúan sobre la misma línea, atendiendo la longitud de las ramas al número de cambios en cada especie y formando un árbol aditivo.

b)En caso de disponer de la secuencia nucleotídica de un gen codificador de proteína para un grupo de especies, ¿Qué información utilizarías preferiblemente para obtener su árbol filogenético y por qué? Haré unos razonamientos que me lleven a algunas conclusiones. Las secuencias de un mismo gen, en un conjunto de especies, serán más distintas cuanto más alejadas filogenéticamente estén las especies comparadas. La acumulación de mutaciones en el ADN a lo largo del tiempo es la causa de que las secuencias de un mismo gen en dos especies distintas no sean idénticas. Cuanto más tiempo pase desde el último antecesor común más diferentes serán las secuencias. Actualmente, los microsatélites son uno de los marcadores más populares en los estudios de caracterización genética. Los microsatélites presentan alta tasa de mutación y naturaleza codominante. Los datos de microsatélites se usan también frecuentemente para evaluar relaciones genéticas entre poblaciones e individuos mediante el cálculo de las distancias genéticas. Podría entonces buscarse una relación entre posibles microsatélites, que suelen darse en intrones del gen.

De otro lado, hemos visto en teoría como se aplican métodos de reconstrucción filogenética en base a las diferencias en las secuencias de nucleótidos de un gen o aminoácidos en una proteína. Estas diferencias se asocian a un ritmo de mutación, que bien puede darnos información del número de sustituciones que han podido haber o, si conocemos datos del reloj molecular en cuestión, proporcionarnos información temporal en cuanto a la especiación. Cabe decir que, al tratarse de especies distintas, los relojes moleculares de un mismo gen o proteína puedan tener distinto ritmo

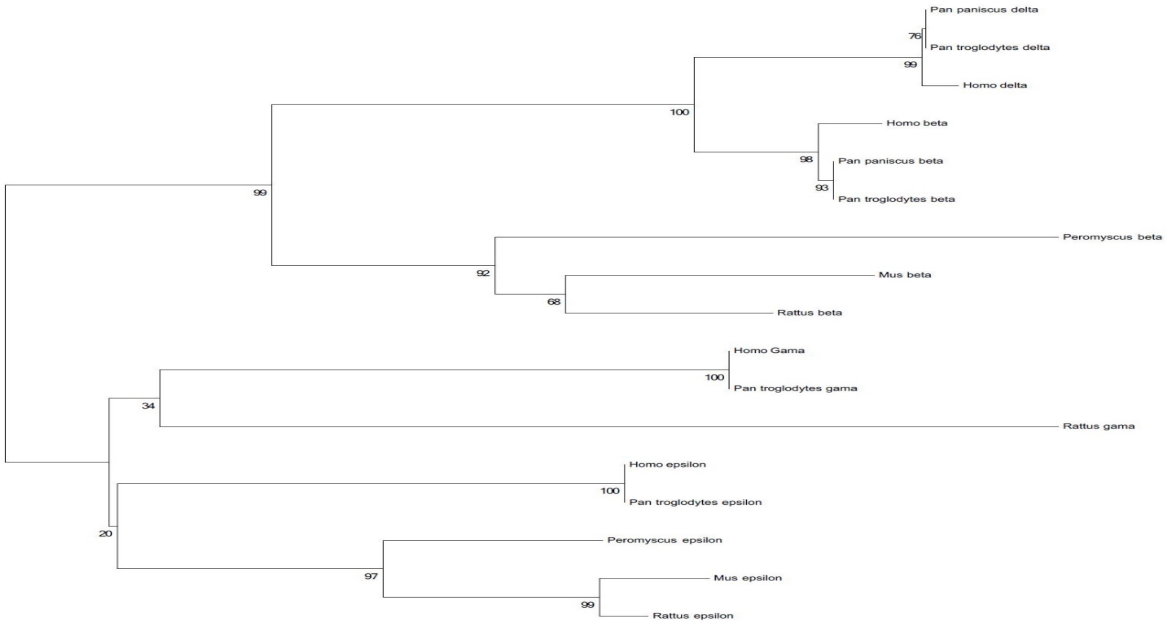
evolutivo, cosa que desconozco. Y de otro lado, las variaciones del ADN se clasifican como "neutras" cuando no originan cambios en los caracteres metabólicos o fenotípicos, y por consiguiente no están sometidas a selección positiva, negativa o de reequilibrio; en caso contrario, se denominan "funcionales". Así pues, dado que sabemos que contamos con especies y un mismo gen ancestral supuesto, convendría fijarse en aquellas mutaciones o cambios en las secuencias que corresponden a regiones codificadoras de proteínas o reguladoras de expresión génica, puesto que es un hecho que el grupo de especies se ha dado. Y de otro lado, cada proteína es un reloj molecular, con lo que, sabiendo si su ritmo de evolución es rápido o lento, podremos aplicarlo para especies que hayan divergido en menor o mayor tiempo.

También deberíamos fijarnos en las limitaciones funcionales de la proteína a analizar, sabiendo que a menor limitación, menos cambios puede acumular y, como es un hecho que existe, los cambios que se hayan dado habrán sido adaptativos de la especie en concreto.

c) Según esta imagen de un árbol de 4 genes (beta delta, épsilon y gama) en diferentes especies (Homo, *Pan paniscus*, *Pan troglodytes*, Peromyscus, Mus, Rattus):

Rellena la tabla siguiente, indicando las relaciones de ortología (O) o de paralogía (P) entre los siguientes pares de secuencias:

Homo delta - Homo èpsilon → Parálogos
Homo delta - Homo beta → Parálogos
Homo beta - Mus beta → Ortólogos
Rattus beta - Mus beta → Ortólogos
Rattus èpsilon - Homo gama → Parálogos
Rattus èpsilon - Peromyscus èpsilon → Parálogos
Rattus gama - Rattus beta → Parálogos
Mus beta – *Pan troglodytes* beta → Ortólogos
Homo beta - *Pantroglodytes* beta → Ortólogos
Mus èpsilon - Mus beta → Parálogos


Pan paniscus delta
76
Pan troglodytes delta
99
Homo delta
100
Homo beta
98
Pan paniscus beta
93
Pan troglodytes beta
99
Peromyscus beta
92
Mus beta
68
Rattus beta
Homo Gama
100
Pan troglodytes gama
34
Rattus gama
Homo epsilon
100
Pan troglodytes epsilon
20
Peromyscus epsilon
97
Mus epsilon
99
Rattus epsilon