PRACTICA 2.

COMPARACIÓN DE SECUENCIAS, EVOLUCIÓN E INFERENCIA FILOGENÉTICA

Genómica Computacional 2019-1
Licenciatura en Ciencias de la Computación
Facultad de Ciencias - UNAM

La práctica deberá ser entregada vía correo electrónico a más tardar el día 15 de octubre a las 23:59 hrs.

Las figuras incluídas deberán llevar un pie de imagen en donde se describa esta de manera concisa. Al final de la práctica se deberá incluir una sección de **referencias** donde se citen las fuentes consultadas que pueden incluir artículos de cualquier tipo, libros de texto o ligas a páginas de internet. Las respuestas en las que se utilice información de alguna fuente revisada deberán incluir la respectiva cita en el formato de su elección.

1. **Identificación de secuencias.** Utilizando BLAST identifica las secuencias que se encuentran en el archivo *pregunta1.fasta*. Para cada una de ellas provee la siguiente información.

    a. Tipo de secuencia.
    b. Longitud de la secuencia.
    c. Organismo asociado a la secuencia con mejor *match*.
    d. Breve descripción del gen/proteína.
    e. Significancia estadística del alineamiento con mejor *match*.

    BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi

2. **Comparación de secuencias e identificación de mutaciones.** Un estudiante hipotético de la Facultad de Ciencias está estudiando neurodesarrollo. A partir de sus estudios encontró una región en el genoma de su organismo de estudio que parece estar involucrada en este fenómeno. Dicho estudiante ha aislado cuatro mutantes de las cuales una muestra un nulo efecto en el fenotipo, otra muestra un leve efecto y las dos muestran un efecto letal en el fenotipo. Utilizando las secuencias del archivo *pregunta2.fasta* ayuda al estudiante hipotético con las siguientes preguntas (en el archivo: WT indica la secuencia de referencia o silvestre y el resto de las secuencias, identificadas como MUTX, son las secuencias mutantes):

    a. ¿Cuál es el organismo que analiza el estudiante hipotético?
    b. Revisando fuentes adicionales, ¿existe evidencia previa de que esta región del genoma esté involucrada en el neurodesarrollo?
    c. ¿Cuántas mutaciones existen entre cada una de las secuencias mutantes respecto a la secuencia silvestre?

d. Para cada una de las mutaciones identificadas en las secuencias mutantes, clasificarlas como (i) sustitución, (ii) deleción o (iii) adición.
e. Para cada una de las secuencias, reporta cada una de las mutaciones de la forma:
    i. En el caso de sustituciones: XYZ, donde X es el nucleótido original, Y es la posición de la mutación y Z el nucleótido producto de la mutación. Por ejemplo: C21T indica que la citosina en la posición 21 fue reemplazada a una timina.
    ii. En el caso de deleciones: ΔXY donde X es el nucleótido original y Y es la posición de la mutación. Por ejemplo ΔC21 indica que se eliminó la citosina en la posición 21.
f. ¿Hay evidencia de que WT sea una secuencia codificante? Justifica tu respuesta. Si es así, ¿a qué gen corresponde?
g. Si la respuesta al inciso (f) es positiva, clasifica las mutaciones identificadas en cada secuencia en (i) sinónimas, (ii) no sinónimas, (iii) corrimiento del marco de lectura y (iv) sin sentido. Si la respuesta al inciso (f) es negativa, deja este inciso sin contestar.
h. Si la respuesta al inciso (f) es positiva, reporta cada una de las mutaciones no sinónimas de la forma: XYZ donde X es el aminoácido original, Y la posición del aminoácido en la proteína y Z el aminoácido producto de la mutación. Ejemplo: A53Y indica que el aminoácido alanina en la posición 53 fue reemplazado por el aminoácido tirosina. Si la respuesta al inciso (f) es negativa, deja este inciso sin contestar.
i. Si la respuesta al inciso (f) es positiva y basado en el tipo de mutación (sinónima, no sinónima, corrimiento del marco de lectura o sin sentido), ¿cuál considerarías que es el efecto de cada una de las secuencias mutantes en el fenotipo (un nulo, un leve y dos letales)? Si la respuesta al inciso (f) es negativa, deja este inciso sin contestar.

3. **Inferencia filogenética y relaciones de parentesco.** El archivo *pregunta3.fasta* contiene las secuencias de la proteína citocromo b obtenida de las mitocondrias para 32 primates y además una de ratón, este último como grupo externo.
    a. Describe brevemente qué es y cuál es la función de la proteína citocromo b.
    b. A partir de consultar fuentes de información adicionales y utilizando los nombres que se encuentren en los encabezados de cada una de las secuencias, clasifica a los organismos para los cuales se proporcionan secuencias en: lemuriformes, tarsiiformes, monos del nuevo mundo, monos del viejo mundo y hominoides.
    c. A partir de consultar fuentes de información adicionales, ¿cuáles es el rasgo característico de los hominoides?
    d. Utilizando IQ-TREE, infiere el árbol filogenético de máxima verosimilitud y visualiza el resultado en FigTree o el paquete ape de R utilizando a *Mus musculus* para enraizar el árbol.
    e. Según el reporte generado por IQ-TREE, ¿cuál es el modelo utilizado para inferir la filogenia?

f. Según la filogenia obtenida, los organismos incluídos en este análisis se agrupan de acuerdo a la clasificación del inciso (*b*).
g. ¿Cuál es el organismo o los organismos más cercanamente relacionado al humano? ¿Cuál es el nombre común de este o estos organismos?
h. ¿Cuál dirías que es el linaje más basal dentro del grupo de los primates?

4. **Integración de filogenias moleculares y rasgos fenotípicos.** Previo a la inferencia filogenética utilizando secuencias moleculares, los caracteres morfológicos eran utilizados para estimar las relaciones de parentesco entre organismos. Aunque los resultados obtenidos a partir de ambos enfoques pueden discernir, considerar ambos es importante para vislumbrar las historias evolutivas. En los archivos *pregunta4_nt.fasta* y *pregunta4_aa.fasta* se encuentran las secuencias de nucleótidos y aminoácidos, respectivamente (las secuencias de aminoácidos corresponden a las secuencias traducidas de las secuencias de nucleótidos), obtenidas de miembros de la familia Volvocaceae. Dentro de esta familia se encuentra el género Volvox, el cual es contiene organismos considerados como multicelulares. Tomando en cuenta los caracteres morfológicos, se ha considerado clásicamente que todos los miembros del género *Volvox* componen un grupo filogenético dentro de la familia Volvocaceae y que por lo tanto la multicelularidad en esta familia ha emergido una única vez. Utilizando las secuencias de estos archivos responde:
    a. Utilizando únicamente una de las secuencias, ¿a qué gen corresponde las secuencias utilizadas en este ejercicio? Describe brevemente qué es y cuál es la función de este gen.
    b. A partir de alinear las secuencias, ¿podrías decir que la secuencia se encuentra conservada a través de los miembros de la familia Volvocaceae?
    c. Dada la respuesta del inciso anterior, ¿qué tipo de secuencia (nucleótidos o aminoácidos) consideras que es más apropiada para inferir una filogenia de este grupo?
    d. Utilizando IQ-TREE y el alineamiento siguiendo la consideración del inciso anterior, infiere el árbol filogenético de máxima verosimilitud y visualiza el resultado en FigTree o el paquete ape de R.
    e. Según el reporte generado por IQ-TREE, ¿cuál es el modelo utilizado para inferir la filogenia?
    f. Considerando la filogenia, ¿se apoya la hipótesis de que la multicelularidad ha emergido en una ocasión dentro de la familia Volvocaceae o de manera se podría proponer de manera alternativa que ha emergido múltiples veces?

5. **Hipótesis de selección con base en secuencias de nucleótidos (opcional).** Utilizando el alineamiento de las secuencias de nucleótidos de las pregunta 4, así como la función *kaks* del paquete seqinr en R. ¿Se podría decir que el gen considerado se encuentra sujeto a selección positiva, selección purificadora o evolución neutral?