PRACTICA 1.

BASES DE DATOS, FORMATOS DE ARCHIVOS Y OPERACIONES FUNDAMENTALES SOBRE SECUENCIAS BIOLÓGICAS

Genómica Computacional 2019-2
Licenciatura en Ciencias de la Computación
Facultad de Ciencias - UNAM

La práctica deberá ser entregada vía correo electrónico a más tardar el día 04 de marzo a las 23:59 hrs

Para esta práctica se revisarán dos paquetes de R: “seqinr” y “Biostrings”.

1. **Lectura de archivos.** Selecciona alguno de los archivos proporcionados en formato fasta. Basado en el archivo de tu elección, responde:
    i. ¿Qué tipo de secuencias se encuentran en cada uno de los archivos el archivo (ADN, ARN y/o proteínas)?
    ii. ¿Cuántas secuencias se encuentran reportadas en el archivo?
    iii. ¿Cuál es la longitud promedio de las secuencias?
    iv. ¿Cuál es la desviación estándar de la longitud de las secuencias en el archivo?
    v. Realiza una gráfica que muestre la distribución de la longitud de secuencias

2. **Contenido de palabras en secuencias.** Utilizando los 5 archivos *Nombre_del_organismo_nt.fasta*
    a. Realiza una gráfica donde se compare la distribución de contenido GC para los diferentes organismos. El contenido GC es una medida que se calcula como (#G + #C)/(longitud de la secuencia).
        i. Realiza una búsqueda de cada uno de los organismos incluidos y describe brevemente a cada uno de ellos.
        ii. ¿Para cuál de los organismos la distribución de contenido GC es más variable?
        iii. ¿Para cuál de los organismos incluidos presenta, en promedio, el mayor contenido GC en sus secuencias? Busca información sobre este organismo y discute cuál podría ser la relación entre un alto contenido de GC y su estilo de vida.
        iv. Para cada colección de secuencias obtén las cadenas inversas complementarias y compara la distribución de GC de estas con las secuencias originales. ¿Cómo se podría explicar esta observación?
3. **Conversión entre tipos de cadenas.** Utilizando el código genético, convierte las cadenas de ADN de los archivos *Nombre_del_organismo_nt.fasta* a cadenas de proteínas.

a. Realiza una gráfica donde se muestre la proporción promedio de cada uno de los aminoácidos en cada uno de los organismos incluídos.
b. ¿Hay algún organismo que muestre una proporción particularmente diferente en alguno de sus aminoácidos respecto a los demás organismos incluídos?
c. ¿Hay algún aminoácido que se encuentre en proporciones similares a través de todos los organismos?

4. **Reconocimiento de patrones y corte de secuencias.** La PCR es una técnica molecular ampliamente usada en biología molecular para aislar y amplificar segmentos específicos de ADN dentro de una molécula más grande. Consiste en identificar dos secuencias (~20 pb) de ADN que puedan usarse para aislar una región deseada. Dada una cadena de ADN, una de las secuencias identificadas debe ser complementaria a una de las cadenas mientras que la segunda secuencia debe ser complementaria a la segunda cadena.

   Considera el siguiente par de secuencias:

   ```
   5′ - CTA CAG CCG TTG CCG AAC GT - 3′
   5′ - AAA AAT ACT CTG CCT TTG AG - 3′
   ```

   a. Asumiendo que la identidad de la base en cada posición es independiente, ¿cuál es la probabilidad de encontrar cada una de las cadenas mostradas en el genoma de *Escherichia coli*?
   b. ¿Cuántas veces se encuentra cada una de las cadenas mostradas en el genoma de *Escherichia coli*?
   c. Asumiendo que el par de secuencias anteriores se utiliza para llevar a cabo la PCR, contesta las siguientes preguntas:
      i. ¿Cuál es la longitud del segmento de ADN que se localiza en la región bordeada por el par de secuencias?
      ii. Utilizando la información del archivo GenBank (https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3), ¿qué elementos de interés se encuentra en esa región?
      iii. Si dentro de la región puede identificarse algún gene, realiza una búsqueda rápida respecto a dicho gen?