Trabajo Práctico (Bioinformática)

Introducción

El presente trabajo práctico consta de tres (3) actividades que se irán introduciendo y desarrollando a lo largo del dictado de la asignatura. El trabajo comenzará con un conjunto de solicitudes que se irán incrementando y refinando a medida que se dicten los temas. Para cada una de las actividades se deberá tener en cuenta la siguiente:

- Objetivos: necesidad a cubrir con el programa.
- Alternativas: las diferentes formas de conseguir los objetivos.
- Desarrollo y Verificación
- Documentación

El desarrollo del trabajo práctico se puede llevar a cabo en forma grupal, pero su evaluación es individual.

Presentación y evaluación

Por cada inciso de cada actividad (1.a, 1.b, etc.) el grupo deberá presentar un archivo .cpp con el programa que resuelve el problema (probado y funcionando). Los archivos deben tener el formato de nombre: "*Act1ayb_Apellido1-Apellido2-Apellido3.cpp*". Es importante documentar el código generado, introduciendo los comentarios correspondientes, de manera que se entienda el razonamiento seguido a la hora de programar.

Planteo del problema

La **secuenciación** es un procedimiento básico en el campo de la investigación en biología y medicina. Gracias a esta técnica, actualmente se ha logrado conocer la molécula esencial de la vida, el ADN, y las 3 mil millones de moléculas que componen nuestra especie (Proyecto Genoma Humano). La secuenciación del ADN permite determinar el orden de los cuatro componentes básicos químicos, llamados "bases", que forman la molécula de ADN. Estas cuatro bases (nucleótidos o aminoácidos) son: Adenina (A), Timina (T), Citosina (C) y Guanina (G). Las mismas forman la doble hélice de ADN uniéndose entre ellas siempre con la misma pareja para formar "pares de bases". La adenina (A) siempre forma pareja con la timina (T); y la citosina (C) se une a la guanina (G). Este emparejamiento es la base para el mecanismo mediante el cual las moléculas de ADN se copian cuando las células se dividen, y también es la base para los métodos usados en la mayoría de los experimentos de secuenciación de ADN.

La información que se almacena en el ADN, se transcribe a ARN, y se traduce luego a proteínas, las cuales son las encargadas de realizar las tareas que están "escritas" en el ADN. Conocer la secuencia de estos pares brinda a los científicos información acerca de los segmentos específicos de ADN, para determinar los tramos que contienen genes y aquellos que transportan instrucciones regulatorias

que activan o desactivan genes; y resaltar los cambios en un gen que pueden causar enfermedades. Para más información puede recurrir a los siguientes enlaces:

https://www.youtube.com/watch?v=MvuYATh7Y74&t=33s

https://es.wikipedia.org/wiki/Secuenciación_del_ADN

http://www.ehu.eus/biofisica/juanma/bioinf/pdf/tema_2a.pdf

https://www.genome.gov/27563183/secuenciacin-del-adn/

La información que se obtiene de la secuenciación del ADN se guarda informáticamente de forma práctica en archivos de formato especiales los cuales se pueden obtener de bases de datos públicas como las de EMBL (Laboratorio Biológico Molecular Europeo), NCBI (Centro Nacional de Información Biotecnológica) o DDJB (Banco Japonés de Datos de ADN):

https://www.ddbj.nig.ac.jp/index-e.html

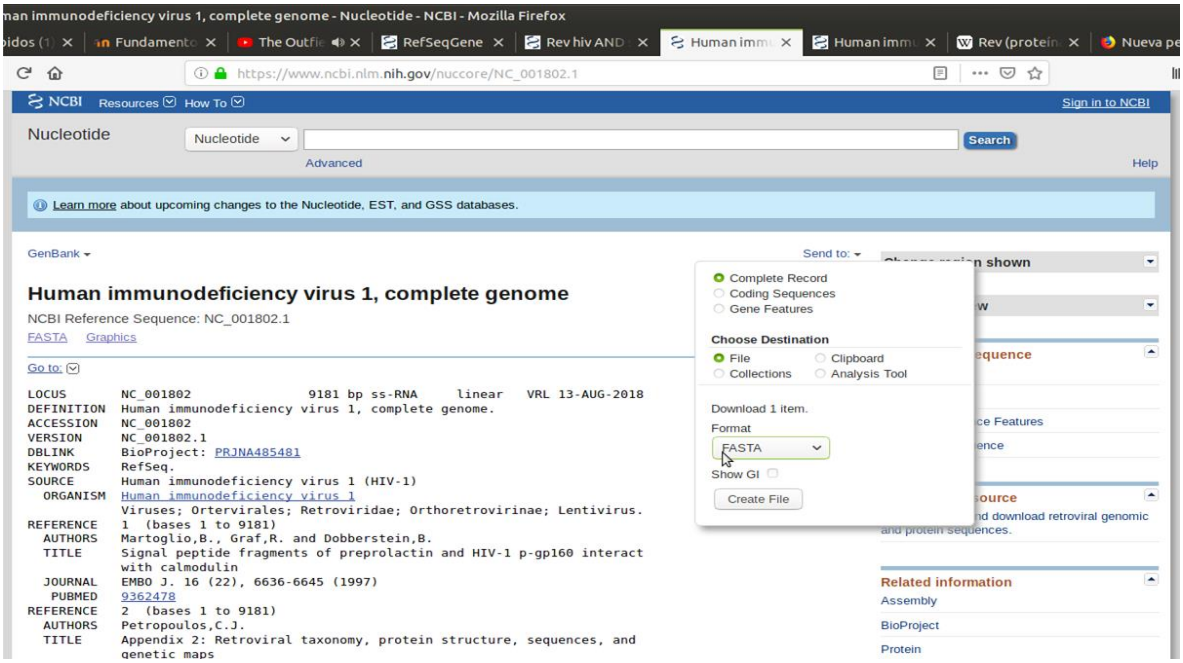
https://www.embl.de/

https://www.ncbi.nlm.nih.gov/.

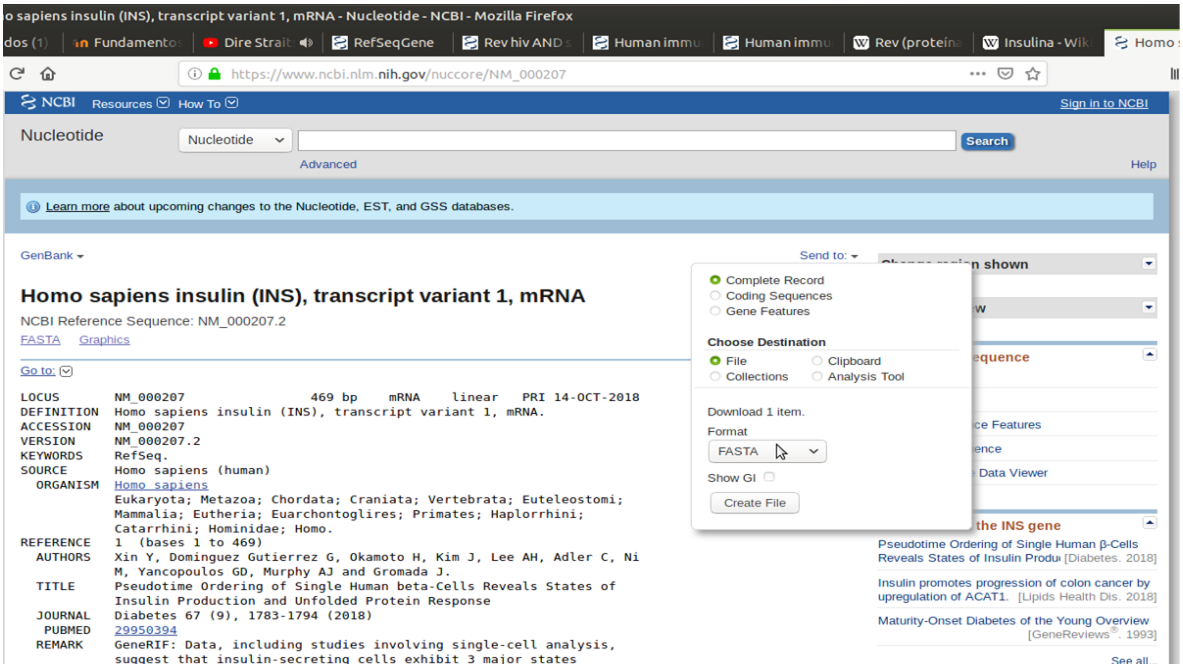
Uno de los objetivos de los Bioinformáticos es poder analizar y obtener conocimiento útil y de calidad a partir de toda esta información disponible, para, por ejemplo, saber si hay algún gen determinado, algún promotor (secuencia encargada de la expresión del gen), si hay presencia de intrones (secuencias que se eliminan en la transcripción) o exones (secuencias que se traducen), o conocer la cantidad de algún nucleótido en particular o en general.

A continuación se muestran ejemplos de 2 archivos con secuencias de ADN y como descargarlos:

- Genoma del VIH-1: https://www.ncbi.nlm.nih.gov/nuccore/NC_001802.1



- Variante 1 del ARNm de la Insulina en Humanos: https://www.ncbi.nlm.nih.gov/nuccore/NM_000207



Actividades

1. Organización de la información y Procesamiento

a) Previamente de ver los archivos de secuencias, genere un arreglo estático de 5000 elementos y escriba en el mismo utilizando la función rand(..) de la biblioteca estándar, los nucleótidos (A,C,T,G).

b) Diseñe e implemente algoritmos para obtener los siguientes parámetros a partir de los datos del arreglo:

i. La cantidad total de C, G, A, T.

ii. La proporción de C y G.

iii. Buscar islas de CpG, indicando la cantidad de las mismas. Se define una isla CpG como una secuencia de 4 o más nucleótidos de C y/o G seguidos (para más información ver https://es.wikipedia.org/wiki/Islas_CpG).

iv. Mostrar la posición y el tamaño de la isla CpG más grande.

c) Diseñe e implemente algoritmos para resolver los siguientes problemas:

i. La PCR (siglas en inglés de Reacción en Cadena de la Polimerasa. Para más información puede ver: https://es.wikipedia.org/wiki/Reacci%C3%B3n_en_cadena_de_la_polimerasa) es la técnica básica para poder a partir de poco ADN obtener una cantidad mucho mayor. La técnica consiste en diseñar pequeños fragmentos de ADN, 15-20 nucleótidos, (también llamados *primers* o cebadores) complementarios a la secuencia a duplicar, que se utilizan para que actúe la ADN Polimerasa (maquinaria celular encargada de duplicar el ADN). La fórmula para calcular la *melting temperature* (temperatura en la cual los *primers* se unen al ADN) es la siguiente:

$$Tm=[4(G+C)+2(A+T)],$$

donde A, C, G, T representan la cantidad de estos nucleótidos.

Diseñe un algoritmo en el cual se ingrese como dato una posición dentro del arreglo, lea a partir de allí en adelante 20 nucleótidos y calcule la fórmula anterior.

ii. Una caja TATA es una secuencias específica de ADN, compuesta por las bases timina y adenina, que se encuentra cerca de algunos genes en arqueas, bacterias y eucariotas. Se estima que el 24% por ciento de los genes humanos contienen la caja TATA. Genere un algoritmo que permita leer una secuencia, como la del arreglo, y que devuelva la cantidad de cajas TATA que presenta la misma.

iii. Codifique un algoritmo que lea una secuencia y calcule su **reversa complementaria**. Debe recordar que el ADN es una molécula que posee una estructura de doble hebra, donde una es **complementaria** a la otra, lo que le da estabilidad a la macro-molécula. Para ello los nucleótidos se complementan de manera tal que la A estabiliza a la T (y viceversa) y el componente G estabiliza a la C. Entonces si tenemos la siguiente secuencia de ADN: ATTATCGCGC. Su reversa complementaria es: TAATAGCGCG, la cual por convención se informa en forma espejada como **GCGCGATAAT** (puede comprobar sus resultados utilizando el siguiente sitio https://www.bioinformatics.org/sms/rev_comp.html).

2. Modularización

a) Proponga funciones que permitan procesar y obtener los resultados que se solicitan en los puntos b y c de la actividad anterior.

b) Implemente una función que reciba una secuencia y determine la cantidad y tamaño y posición de las islas CpG. La función debe devolver la cantidad islas CpG encontradas, la longitud (cantidad de nucleótidos) de la mayor de ellas y su posición.

c) (**Opcional)** *Implemente una función que permita determinar si es posible encontrar la presencia de un gen dentro de una secuencia dada, a partir de las posiciones de las islas de CpG y posiciones de la caja TATA. Para ello debe poder determinar los lugares en los que se encuentran islas CpG y cajas TATA, y si las diferencias entre ambas posiciones están en un rango de 70 a 90 nucleótidos, estamos en presencia de un posible gen eucarionte. La función debe retornar un valor indicando la cantidad de posibles genes detectados.*

3. Almacenamiento

a) Utilice los archivos FragmentoCromo1.txt y FragmentoCromo1.hea los cuales contienen un fragmento de ADN del Cromosoma 1 humano, para obtener y mostrar los datos que se solicitan en la actividad anterior.

b) Utilice ahora el archivo FragmentoCromo1.fasta para obtener la información que se solicita en el punto anterior. En bioinformática, el formato FASTA se utiliza para representar secuencias de ácidos nucleicos y péptidos. Este formato se basa en texto y los pares de bases o los aminoácidos se representan usando códigos de una única letra. El formato también permite incluir nombres de secuencias y comentarios que preceden a las secuencias en sí.

c) Genere una función llamada RelaciónCG (...) que obtenga, a partir de la secuencia del archivo, una lista con las posiciones de las islas CpG, el tamaño de cada y su relación con la secuencia completa.

d) Genere, por cada punto que sigue, una archivo de texto donde se guarde la siguiente información:

i. La cantidad total de cajas TATA y la posición que ocupa cada una.

ii. La cantidad de islas CpG, la posición que ocupa cada una y su longitud.

iii. A partir de los resultados de la función RelaciónCG (...) del punto c, genere la siguiente salida:

Posición	Tamaño	Relación
150	100	0.01
1000	20	0.002
…	…	…

iv. G, A, T. La proporción de C y G.

e) Guarde en formato binario la información de los archivos trabajados anteriormente.