Tarea 2 - Programación Funcional

INS125 - Leguajes de programación
Universidad Andrés Bello

28 de mayo de 2021

1. Situación

Dada la situación sanitaria mundial, usted ha decidido poner a disposición del pais sus conocimientos y ha montado un laboratorio de toma y secuenciación de test PCR. Por esta razón, una vez que obtenida la muestra desde el sujeto, usted se dedicará a secuenciar y reportar los datos obtenidos.

2. Positividad a SARS-CoV-2

La extracción de la muestra y luego la obtención de la secuencia del virus serán utilizadas para determinar si es un caso positivo de SARS-CoV-2 y detectar a que cepa pertenece.

Una secuencia del ADN del virus corresponde a una cadena de 66 caracteres compuestas por las letras A, T, C y G, los cuales representa los compuestos químicos adenina (A), timina (T), citosina (C) y guanina (G), respectivamente; que constituyen las bases de nucleótidos del ADN.

Una muestra extraída a una persona es positiva para SARS-CoV-2 si su valor para la proteína Spike es mayor o igual a 8, siendo $A = 1, T = 2; C = 3; G = 4$. La proteína Spike está compuesta de una secuencia de 4 compuestos ubicados en la posición 40 a 43. En la figura 1 se observa una muestra muestra positiva y otra negativa para SARS-CoV-2.

3. Secuenciación y detección de cepas

Mundialmente existen cerca de 2 mil cepas diferentes de coronavirus. En nuestro laboratorio, tenemos especial interés en detectar las cepas B.1.1.7, B.1.351 y P.1, principalmente debido a que tenemos recursos limitados para generar la secuencia de ADN y solo podemos sintetizar 66 de los mas de un millón de caracteres que componen el ADN del virus.

Cada una de las cepas tiene distinto método de detección, los cuales se detallan a continuación.

Figura 1: Ejemplo de una secuencia de ADN positiva y negativa para el virus SARS-CoV-2

	A	G	C	T	C	C	A	A	G	G	T	
T	2											
G		16										
G			12									
A				2								
A					3							
C						9						
C							3					
T								2				
C									12			
G										16		
A											2	
												79

Figura 2: Ejemplo de detección para la cepa P.1 en una simplificación de secuencia de ADN (simplificada a solo 10 caracteres)

- Para detectar si la secuencia corresponde a la cepa B.1.1.7, se debe verificar que la cantidad de compuestos T dividido la cantidad de compuestos A sea mayor a 1.
- Para detectar si corresponde a B.1.351, el valor de la suma de sus compuestos pares mas el cero es mayor que el de sus compuestos impares (considerando $A = 1, T = 2; C = 3; G = 4$ y que el primer índice es cero)
- Para detectar la Cepa P.1, siendo esta la cepa mas difícil de detectar, se debe establecer una matriz, donde las filas corresponderán a los 66 elementos de la secuencia, y las columnas corresponderán a los mismo 66 elementos en orden inverso. Luego, usted debe obtener la diagonal de esa matriz construida en base a la multiplicación de su fila-columna, y sumar sus componentes. Si el resultado es mayor a 525, se considera positiva a dicha cepa. En la figura 2 usted podrá observar un ejemplo para una secuencia de solo 10 caracteres (por motivos pedagógicos).
- Una secuencia de ADN solo puede pertenecer a una Cepa. Entonces, la supremacía en la detección viene dado en orden inverso a como se presenta en este enunciado. Por ejemplo, Si haciendo los cálculos correspondientes a la detección, y arroja positivo para B.1.1.7 y también para P.1, el resultado debe ser que esa muestra pertenece a P.1.

4. Tarea

Comprendiendo lo anterior, y considerando que desde laboratorio de toma de muestras, usted recibirá un archivo TSV (archivo de texto separado por tabulaciones) llamado **examenes.tsv** cual contiene nombre, rut y secuencia de ADN de la muestra, se le solicita realizar lo indicado siguiendo las siguientes indicaciones:

1. Filtre los exámenes e incorpore los resultados a los archivos **positivos.tsv** y **negativos.tsv** los cuales contendrán los exámenes que dieron positivo y negativo a SARS-CoV-2, respectivamente.
2. Aplique la función $conversionNumerica(string)$ sobre todos los exámenes y escriba un archivo llamado **numericos.tsv** con los datos de la persona y su correspondiente secuencia en formato numérico, donde $A = 1, T = 2; C = 3; G = 4$.
3. Filtre los exámenes que **dieron positivo a SARS-CoV-2** y genere un archivo de texto por cada una de las cepas detectadas, **B117.tsv**, **B1351.tsv** y **P1.tsv** el cual contendrá los datos de las personas que dieron positivo a dichas cepas.
4. Produzca un archivo **logs.txt** el cual contenga la cantidad de exámenes analizados, la cantidad de positivos, los que pertenecen a la cepa B.1.1.7, B.1.351 y P.1; un dato por cada linea.

5. Restricciones

- Debe implementar las funciones (como funciones con identificador):
  - $leerArchivo(nombrearchivo)$
  - $conversionNumerica(string)$
  - $detectarCepaB117(string)$
  - $detectarCepaB1351(string)$
  - $detectarCepaP1(string)$

  Todas las demás funciones que utilice su programa deben ser funciones anónimas (*Lambda*) utilizando el paradigma funcional.
- Para desarrollar su solución deberá utilizar las funciones map, reduce y filter.
- No está permitido hacer más imports de los ya existentes. Es decir, sólo puede utilizar las bibliotecas `string`, `functools`

6. Instrucciones

- Fecha de entrega: Domingo 6 de Junio, 2021 a las 23:59.
- Trabajo personal hecho en lenguaje Python.
- Método de entrega: Su repositorio privado creado a través del link de la tarea. **No se aceptaran entregas en repositorios no creados a través del link de la tarea.**
- Para comenzar su tarea, clone su repositorio y utilice el archivo `main.py`, con la estructura del programa.
- Su repositorio de la tarea debe contener únicamente su archivo `main.py`. Si durante su desarrollo trabaja con otros archivos, Utilice un archivo `.gitgnore` para no subirlos a su repositorio remoto.
- Su programa debe recibir y generar los archivos entrada y salida indicados en el enunciado.
- El proceso de revisión será automatizado. Es importante respetar el **formato establecido para los archivos de entrada y salida. Este formato no es modificable**. Por lo tanto, si su archivo de salida no corresponde al formato preestablecido o no existe, su calificación será mínima.
- Las preguntas sobre la tarea deben ser formuladas como un Issue en el repositorio del laboratorio ubicado en el siguiente link https://github.com/INS125/Laboratorio/issues

7. Recomendaciones

- Se recomienda hacer commits parciales. Si sus archivo de salida no contiene toda la matriz o el resultado completo de la Convolución pero si contiene parcialmente la matriz, puede obtener puntaje parcial.
- Si su programa no crea el o los archivos de salida, o no se ejecuta, será evaluado con la nota mínima.
- En el repositorio oficial del laboratorio puede encontrar dos ejemplos de archivos de entrada y su correspondiente salida. https://github.com/INS125/Laboratorio/
- Recuerde solicitar unirse a github student. Si no lo hace, no podrá hacer su código privado y cualquiera podría visualizar su tarea.
- Es un trabajo personal. Es su responsabilidad cuidar su tarea.

8. Código de honor

Toda persona inscrita en este curso se compromete a:

- Actuar con honestidad, rectitud y buena fe frente a sus profesores y compañeros.
- No presentar trabajos o citas de otras personas como propias o sin su correspondiente citación, ya sea de algún compañero, libro o extraídos de internet como también a no reutilizar trabajos presentados en semestres anteriores como trabajos originales.
- No copiar a compañeros ni hacer uso de ayudas o comunicaciones fuera de lo permitido durante las evaluaciones.

Cualquier alumno o alumna que no respete el código de honor durante una evaluación (sea este la entrega de una tarea o el desarrollo de una prueba o control tanto durante la cátedra como el laboratorio) será evaluado con la nota mínima y será virtud de profesor, de acuerdo con la gravedad de la falta, las acciones siguientes a tomar.