Tarea 1

Genómia Computacional / 2023-I

29 de septiembre de 2022

1. Tutorial de expresiones regulares

El uso de expresiones regulares (regex) en Python se hace a través del módulo `re`.

```
>>> import re
```

De este módulo hay que usar las funciones `search` que recibe un patrón a buscar (expresión regular) y la cadena donde se hace la búsqueda

```
>>> re.search(r'TA','GATTACA')
<_sre.SRE_Match object; span=(3, 5), match='TA'>
>>> re.search(r'TA','gattaca', flags=re.IGNORECASE)
<_sre.SRE_Match object; span=(3, 5), match='ta'>
r'AT','GATTACA
>>>
```

Mientras que la función `match` es que busca el patrón al inicio de la cadena y si no lo encuentra entonces no regresa nada.

```
>>> re.match(r'GAT+A','GATATACA')
<_sre.SRE_Match object; span=(0, 4), match='GATA'>
>>> re.match(r'GAT+A','CGATATACA')
>>>
```

También es posible diseñar una expresión regular que pueda definir grupos un diccionario con llaves predifindas.

```
>>> cadenas = ['sec1:ATGAT', 'sec2:CGGANG']
>>> patron = r'(?P<titulo>\w+):(?P<secuencia>[ACGT]*)'
>>> [re.search(patron,cad).groupdict()['secuencia'] for cad in cadenas]
['ATGAT', 'CGGA']
>>> [re.search(patron,cad).groupdict()['titulo'] for cad in cadenas]
['sec1', 'sec2']
```

En el ejemplo siguiente se definen dos grupos: `titulo`, definido por más o un símbolo alfanumérico (`\w+`) y el otro grupo es `secuencia` definido por 0+ símbolos de `A,C,G,T`. Los resultados se agrupan con una lista de comprensión y se muestran las apariciones encontradas.

2. Instrucciones

Responde las siguientes preguntas y haz las implementaciones en código como se indica a continuación:

- Las respuestas que lo requieran, escríbelas en un archivo pdf en el que venga una lista de los integrantes del equipo en orden alfabético
- Haz las implementaciones de los problemas que lo requieran en un script de PYTHON que se llame `tarea_n.py`.
- Sube todos los archivos a la entrada de la tarea correspondiente en el classroom (sin comprimir).

3. Ejercicios

1. Realiza los siguientes ejercicios sobre expresiones regulares.

   a) Construye una expresión regular que reconozca `GGACC` o `GGTCC`

   b) Tomando en cuenta la siguiente expresión regular, `r'G[AC](T*|AC)GG'`, escoge las cadenas que la contienen. También explicala en lenguaje natural.

      1) CTTGG
      2) GCTTGG
      3) GCACGG
      4) GCAACGG

   c) Usando las funciones `re.compile` y `re.finditer` reporta las posicoines en la cadena `GATTATATACATAGTAGTATA` donde se encuentra la expresión regular `r'(TA)+'`. Explica la expresión regular.

   d) ¿Qué cadenas se describen con la expresión regular `r'G.*G'`? Da un par de ejemplos (El alfabeto es $\Sigma = \{A, C, G, T\}$).

2. A continuación se presentan 10 secuencias hipotéticas. Diseña una expresión regular que detecte genes y específica cuales de ellos la cumplen, asumiendo que estos se component por uno de tres codones de inicio, uno y solo un codón de paro que indica el final del gen, mientras que al interior puede haber una cantidad de 1 o más tripletes de nucleótidos.

```
0. ATATATACATACTGGTAATGGGCGCGCGTGTGTTAAGTTCTGTTGTAGGGGTGATTAGGGGCG
1. GGCCCACACCCCACACCAATATATGTGGTGTGGGCTCCACTCTCTCGCGCTCGCGCTGGGGAT
2. ATAAGGTGTGTGGGCGCGCCCCGCGCGCGCGTTTTTTCGCGCGCCCCCGCGCGCGCGCGCGCG
3. GGCGCGGGACGCGGCGGCGGATCCCGATCCGTGCGTCAATACTATTATGGCCAGATAGAATAA
4. GTGCTGCTGCGGCGCCCACACCTATTATCTCTCTCTCTCTGCCTCTCCACCTCGGGGCTTAAT
5. GCGCTGCTGCTGGCTCGATGGGCGCGTGCGTCGTAGCTCGATGCTGGCTCGAGCTGTAATCTT
6. GGCGCTCGCTCGGATGCGCGGCCGGGCTCTCTGCTCGCGCTCGCTTCGCGCTCGTGACCGCTG
7. AATTGGTGCGCGCTCGCGCACACACAGAGAGAGGGTTTATATAGGATGATATATCCACATTGG
```

```
8. ATGCTGCTGCTGGCTCTGCTTGCGCTCTGCTCGCTGGGGTGTGTGTGCCGCGCGCTGCTGCTC
9. GCTGGGCTCGCTCGATGCGCGCGGGCGCGCGACCGCGGACGGCGTCGCTGCTAAATGGGCTTC
```

Debes entregar una lista con el índice de la línea en las que hay un gen válido, ejemplo, si en las líneas 0, 1 y 5 hay, entonces el resultado debería ser:

$$[0, 1, 5]$$

3. En el archivo `promotores.txt` se encuentra la lista de secuencias tomadas del genoma de *Vitis vinifera* y cada una de las secuencias puede que tenga alguno de las diferentes formas en las que se ha encontrado el promotor GATA:

   $\{AGATAG, TGATAG, AGATAA, TGATAA\}$

   Deseamos estudiar estas regiones en función del promotor GATA y por lo tanto lo primero que deseamos es saber cuántas veces aparecen los promotores en cada región.

   Tu respuesta deberá ser el conteo de la cantidad de veces que la región i-ésima contiene al promotor GATA en cualquiera de sus formas. Ejemplo:

   $$[3, 2, 1, ...]$$

   Quería decir que la primer secuencia contiene 3 instancias del promotor, la segunda tiene 2, la tercera 1 y así en adelante.

4. Haz un script en python que reciba como parámetro un entero que determinará la cantidad de iteraciones para el siguiente algoritmo

   Data: M iteraciones
   Result: Número flotante x
   $i \leftarrow 1$;
   $D \leftarrow 0$;
   while $i < M$ **do**
   $i \leftarrow i + 1$;
   $x \leftarrow -1 \leq uniform() \leq 1$;
   $y \leftarrow -1 \leq uniform() \leq 1$;
   $d \leftarrow \sqrt{x^2 + y^2}$;
   if $d \leq 1$ **then**
   $D \leftarrow D + 1$;
   end
   end
   $x \leftarrow 4 \cdot D/i$;
   return x;

   Algorithm 1: Algoritmo misterio

Tip: Revisa la documentación del paquete `random`, particularmente de la biblioteca de `numpy`.

¿Qué está calculando el algoritmo 1?

5. Utilizando los siguientes datos de sensibilidad (93 por ciento) y especificidad (99 por ciento) reportados para cierta prueba rápida de antígeno para detectar la infección por virus SARS-COV2 y considerando una prevalencia actual de COVID en México estimada a partir del promedio de casos nuevos observados a lo largo de 2 semanas de 16000 casos activos respecto a una población total de 120000000 de habitantes encuentra:

   a) ¿Cuál es la probabilidad de que si uno de ustedes se realiza una prueba rápida de este tipo y ésta resulta positiva ustedes en realidad sean portadores del virus SARS-COV2?

   b) ¿Cuál es la probabilidad de que si la prueba resulta negativa ustedes en realidad no sean portadores del virus SARS-COV2?

   c) Entre marzo y junio de 2021 se tuvo un promedio de nuevos contagios semanales de alrededor de 3000 casos, por lo que a lo largo de dos semanas se tendría una prevalencia aproximada de 6000 casos activos respecto a 120000000 de habitantes. Calcula las probabilidades referidas en los dos incisos anteriores pero considerando este nuevo dato de prevalencia. ¿Qué puedes concluir respecto a las probabilidades obtenidas en ambos escenarios?, ¿consideras que en el caso de las pruebas de detección de COVID es necesaria una mayor sensibilidad o una mayor especificidad? Justifica tu respuesta.