Tarea 1

Genómica Computacional - 2019-2

Entrega: Viernes 8 de marzo

El uso de expresiones regulares (regex) en `Python` se hace a través del módulo `re`. Esto es, ya sea en el intérprete o -preferentemente- al inicio de un *script*, hay que importar el módulo mencionado:

```
>>> 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
>>>
```

La diferencia con 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')
>>>
```

1. Realiza los siguientes ejercicios sobre **Expresiones Regulares** con la sintaxis de `Python`.

   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) La sintaxis `R{n,m}` (donde R es una expresión regular válida) significa que R debe aparecer entre `n` y `m` veces. Si buscásemos la expresión `r'T{3,5}'` en la cadena `'GATTATTTACTTTTACAGGT'` ¿en qué posiciones reportaría `Python` encontrar dicha expresión regular.

   e) ¿Qué cadenas se describen con la expresión regular `r'G.*G'`? Da un par de ejemplos.

2. Recordando que la definición de **entropía de Shannon** de la variable aleatoria $X = x_1, x_2, x_3, \ldots, x_n$ es como sigue

$$H(X) = -\Sigma p(x_i) \cdot \log_2(p(x_i))$$

   responde las siguientes preguntas.

   a) ¿Cuál es la entropía de Shannon en bits de las siguientes cadenas? En cada caso reporta la distribución de los símbolos que las conforman

      1) 00101111010101111000000101011
      2) 10010110011110111011011100111
      3) 00000000000000000000000000000

   b) ¿Cuál es la entropía de la cadena $G = ACC\,CTC\,GGC\,GTC$?

   c) Reproduce la gráfica de un proceso Bernoulli. Para esta pregunta describe el proceso matemático o el código para tal reproducción. ¿Cuál es su interpretación?

Deberás entregar un PDF con cada respuesta señalada con el número correspondiente (ej: 1.a.1), el código debera estár en un script de `Python` y el título el correo debera ser:
Tarea1 - *número de cuenta*.