Universidad Nacional Autónoma de México

Facultad de Ciencias

Genómica Computacional

Tarea 1

Expresiones Regulares

Profesor:

Sergio Hernández López

Integrantes:

Guzmán Cortés Dulce Dyliang

Marín Parra José Guadalupe de Jesús

Rangel Limón Erik

1. Realiza los siguientes ejercicios sobre expresiones regulares.

   a) Construye una expresión regular que reconozca GGACC o GGTCC.
   RESPUESTA.
   La expresión regular en tal caso es r'(GG)A*T*(CC)'

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

   1) CTTGG
   RESPUESTA.
   No pertenece ya que esta cadena particularmente inicia con 'C' por lo que la expresión no la acepta.

   2) GCTTGG
   RESPUESTA.
   Si la acepta ya que la cadena inicia con 'G', después vamos con una 'C' para cumplir con [AC], después una cantidad arbitraria de 'T' para cumplir con T* y finalmente termina la cadena con 'GG' como en la expresión regular.

   3) GCACGG
   RESPUESTA.
   Si la acepta ya que la cadena inicia con 'G', después vamos con una 'C' para cumplir con [AC], después con 'AC' para cumplir con (AC) y finalmente termina la cadena con 'GG' como en la expresión regular.

   4) GCAACGG
   RESPUESTA.
   No pertenece a la expresión ya que la rompe en 'AA', esta subcadena no existe dentro de la expresión.

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

```python
#Ejercicio 1 (c)
cadena = 'GATTATATACATAGTAGTATA'
er = 'TA'
print('Dada la cadena',cadena)
for match in re.finditer(er, cadena):
    s = match.start()
    e = match.end()
    print('Extiste un "%s" en %d:%d' % (cadena[s:e], s, e))
```

```
→  Solucion git:(main) ✗ python3 tarea_1.py
Dada la cadena GATTATATACATAGTAGTATA
Extiste un "TA" en 3:5
Extiste un "TA" en 5:7
Extiste un "TA" en 7:9
Extiste un "TA" en 11:13
Extiste un "TA" en 14:16
Extiste un "TA" en 17:19
Extiste un "TA" en 19:21
```

Código y ejecución de las posiciones de la expresión regular en la cadena utilizando python.

La expresión regular detectará al menos una subcadena 'TA' dentro de una cadena, razón por la que en la cadena propocionada existen 7 ocurrencias de 'TA'.

d) ¿Qué cadenas se describen con la expresión regular r'G.*G'? Da un par de ejemplos (El alfabeto es $\Sigma = \{A, C, G, T\}$).
RESPUESTA.
Las cadenas que acepta la expresión regular dada son todas aquellas que inicien y terminen con 'G' sin importar los caractéres que haya entre éstas. Tenemos los siguientes ejemplos.

- GATTATATACATAGTAG y GACGTGTCAG son aceptadas porque inician y terminan con 'G'.

2. Implementación en el archivo tarea_1.py

3. Implementación en el archivo tarea_1.py

4. Implementación en el archivo tarea_1.py. ¿Qué está calculando el algoritmo 1?
RESPUESTA.
Un número según las iteraciones y los cálculos dentro del algoritmo que está condicionado al número ingresado como parámetro.

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 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?
RESPUESTA.

Sea P el evento en donde una persona sea portadora del virus SARS-COV2.
Esta tiene una probabilidad de

$$\mathbb{P}(P) = \frac{16000}{120000000} = \frac{16}{120000} = \frac{8}{60000} = \frac{4}{30000} = \frac{2}{15000} = \frac{1}{7500}$$

La sensibilidad de la prueba nos dice la probabilidad con la que un individuo realmente contagiado sea detectado como positivo, es decir, si S es el evento en donde la prueba sale positiva, entonces la sensibilidad es:
$\mathbb{P}(S|P) = \frac{93}{100}$
Y lo que nosotros queremos calcular es la probabilidad de que un individuo sea portador dado que la prueba salió positiva, es decir, calcularemos la probabilidad del evento $P|S$.
Por la definición de probabilidad condicional sabemos que

$$\mathbb{P}(S|P) = \frac{\mathbb{P}(S \cap P)}{\mathbb{P}(P)}$$

Por lo que

$$\mathbb{P}(S \cap P) = \mathbb{P}(S|P) \cdot \mathbb{P}(P) = \frac{93}{100} \cdot \frac{1}{7500} = \frac{93}{750000}$$

Así mismo la especificidad es la probabilidad de que la prueba salga negativa dado que el individuo no es portador, por lo que tenemos.

$$\mathbb{P}(S^c|P^c) = \frac{99}{100}$$

Y así mismo

$$\mathbb{P}(S|P^c) = \frac{1}{100}$$

Y por la definición de probabilidad condicional tenemos que

$$\mathbb{P}(S|P^c) = \frac{\mathbb{P}(S \cap P^c)}{\mathbb{P}(P^c)}$$

Por lo que

$$\mathbb{P}(S \cap P^c) = \mathbb{P}(S|P^c) \cdot \mathbb{P}(P^c) = \frac{1}{100} \cdot \frac{7499}{7500} = \frac{7499}{750000}$$

Por teoría de conjuntos sabemos que

$$S = (S \cap P) \cup (S \cap P^c)$$

Por lo que, como $(S \cap P) \cap (S \cap P^c) = \emptyset$, seguimos entonces que

$$\mathbb{P}(S) = \mathbb{P}(S \cap P) + \mathbb{P}(S \cap P^c) = \frac{93}{750000} + \frac{7499}{750000} = \frac{7592}{750000} = \frac{949}{93750}$$

Conociendo estos valores podemos entonces calcular la probabilidad condicional del pevento $P|S$, el cual es.

$$\mathbb{P}(P|S) = \frac{\mathbb{P}(S \cap P)}{\mathbb{P}(S)}$$

$$\mathbb{P}(P|S) = \frac{\frac{93}{750000}}{\frac{949}{93750}} = \frac{93}{7592} \approx 0.0122497$$

Por lo tanto hay una probabilidad del $1.22497\,\%$ de que sea portador.

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

Ésta es la probabilidad entonces de que suceda el evento $P^c|S^c$
Su probabilidad es

$$\mathbb{P}(P^c|S^c) = \frac{\mathbb{P}(P^c \cap S^c)}{\mathbb{P}(S^c)}$$

Obtendremos la probabilidad del evento $P^c \cap S^c$.

Ya conocemos la probabilidad del evento $S^c|P^c$, pues ésta es la especificidad de la prueba, el cual es

$$\mathbb{P}(S^c|P^c) = \frac{99}{100}$$

Y por la definición de probabilidad condicional sabemos que esto e

$$\mathbb{P}(S^c|P^c) = \frac{\mathbb{P}(S^c \cap P^c)}{\mathbb{P}(P^c)}$$

Por lo tanto

$$\mathbb{P}(S^c \cap P^c) = \mathbb{P}(S^c|P^c) \cdot \mathbb{P}(P^c)$$

$$\mathbb{P}(S^c \cap P^c) = \frac{99}{100} \cdot \frac{7499}{7500} = \frac{247467}{250000}$$

Por otra parte, ya calculamos la probabilidad de S, la cual fue

$$\mathbb{P}(S) = \frac{949}{93750}$$

Por lo que la probabilidad de su complemento es

$$P(S^c) = \frac{92801}{93750}$$

De manera que

$$\mathbb{P}(P^c|S^c) = \frac{\frac{247467}{250000}}{\frac{92801}{93750}} = \frac{742401}{742408} \approx 0.999991$$

Por lo que la probabilidad es de 99.9 % de que una persona no esté contagiada si la prueba fue negativa.

c) Entre Marzo y Junio de 2021 se tuvo un promedio de nuevos contagios semanales al rededor 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 puede 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.
RESPUESTA.

En este caso entonces la probabilidad de estar contagiado sería de

$$\mathbb{P}(P) = \frac{6000}{120000000} = \frac{1}{20000}$$

Por lo que la probabilidad del primer inciso tendríamos lo siguiente.

$$\mathbb{P}(S|P) = \frac{\mathbb{P}(S \cap P)}{\mathbb{P}(P)}$$

Por lo que el evento $S \cap P$ tiene una probabilidad de

$$\mathbb{P}(S \cap P) = \mathbb{P}(S|P) \cdot \mathbb{P}(P)$$

$$\mathbb{P}(S \cap P) = \frac{93}{100} \cdot \frac{1}{20000} = \frac{93}{2000000}$$

El evento $S^c|P^c$ es la especificidad de la prueba, la cual tiene una probabilidad de 99 %. Por lo que la probabilidad del complemento sería

$$\mathbb{P}(S|P^c) = \frac{1}{100}$$

Así mismo la definición de esta probabilidad condicional es la siguiente.

$$\mathbb{P}(S|P^c) = \frac{\mathbb{P}(S \cap P^c)}{\mathbb{P}(P^c)}$$

Por lo que la probabilidad del evento $S \cap P^c$ es la siguiente

$$\mathbb{P}(S \cap P^c) = \mathbb{P}(S|P^c) \cdot \mathbb{P}(P^c)$$

$$\mathbb{P}(S \cap P^c) = \frac{1}{100} \cdot \frac{19999}{20000} = \frac{19999}{2000000}$$

Por teoría de conjuntos sabemos lo siguiente

$$S = (S \cap P) \cup (S \cap P^c)$$

Y como son disjuntos, seguimos entonces que

$$\mathbb{P}(S) = \mathbb{P}(S \cap P) + \mathbb{P}(S \cap P^c)$$

$$\mathbb{P}(S) = \frac{93}{2000000} + \frac{19999}{2000000} = \frac{20092}{2000000} = \frac{5023}{500000}$$

Por lo tanto el evento $P|S$ tiene una probabilidad de

$$\mathbb{P}(P|S) = \frac{\mathbb{P}(P \cap S)}{\mathbb{P}(S)}$$

$$\mathbb{P}(P|S) = \frac{\frac{93}{2000000}}{\frac{5023}{500000}} = \frac{93}{20092} \approx 0.00462871$$

Por lo que la probabilidad de que si nuestra prueba salga positiva y en realidad seamos portadores es de 0.462871 %.
Por otro lado, con respecto al segundo inciso, seguimos que

$$\mathbb{P}(S^c|P^c) = \frac{99}{100}$$

$$\mathbb{P}(S^c|P^c) = \frac{\mathbb{P}(S^c \cap P^c)}{\mathbb{P}(P^c)}$$

Por lo tanto la probabilidad del evento $S^c \cap P^c$ es la siguiente

$$\mathbb{P}(S^c \cap P^c) = \mathbb{P}(S^c|P^c) \cdot \mathbb{P}(P^c)$$

$$\mathbb{P}(S^c \cap P^c) = \frac{99}{100} \cdot \frac{19999}{20000} = \frac{1979901}{2000000}$$

Y por otro lado ya calculamos la probabilidad de S la cual fue de $\frac{5023}{500000}$, por lo que la probabilidad de su complemento es de

$$\mathbb{P}(S^c) = 1 - \mathbb{P}(S) = 1 - \frac{5023}{500000} = \frac{494977}{500000}$$

Por lo tanto el evento $P^c|S^c$ tiene una probabilidad de

$$\mathbb{P}(P^c|S^c) = \frac{\mathbb{P}(S^c \cap P^c)}{\mathbb{P}(S^c)}$$

$$\mathbb{P}(P^c|S^c) = \frac{\frac{1979901}{2000000}}{\frac{494977}{500000}} = \frac{282843}{282844} \approx 0.999996$$

Por lo que la probabilidad de que realmente no sea portador un individuo si su prueba salió negativa es de 99.999996 %.

En el caso particular de México se necesitaría una prueba con mejor sensibilidad, pues la probabilidad de que un individuo esté realmente contagiado cuando su prueba salió positiva es mínima.

Notemos que esto principalmente se debe a dos cosas; que con tan pocos contagios disminuye considerablemente la probabilidad de que el resultado sea positivo (sin importar si es cierto), el cual es el evento S; y debido a esto disminuye aún más la probabilidad de que suceda que la prueba sea positiva y el individuo sea realmente portador del virus SARS-COV2, el cual también es una probabilidad mínima, (el evento $S \cap P$), como ambas son determinantes al momento de calcular la probabilidad del evento $P|S$, por eso se tiene una probabilidad tan baja.

No es tan necesaria una mejor especificidad pues es mucho mayor la proporción de la gente que no es portadora del virus con respecto a la población total, lo que hace que un resultado negativo sea fiable, pues en el cálculo de dicha probabilidad influye esta proporción.