TAREA2

HERRAMIENTAS BIOINFORMÁTICAS PARA LAS CIENCIAS BIOLÓGICAS:
INTRODUCCIÓN A PYTHON

Introducción

Se sabe que diversas proteínas de unión a DNA se unen a la caja CCAAT, un elemento que actúa en cis encontrado en los promtores de un gran número de organismos. Estas proteínas son conocidas como factor de unión a CCAAT ("*CCAAT-binding factor: CBF*") o NF-Y [1] [2]. Las proteínas CBF son activadores transcripcionales responsables de la activación de muchos genes implicados en el metabolismo [3]. El complejo Hap en levaduras activa al gen de citocromo C (*CYC1*) así como otros genes implicados en el transporte de electrones [4].

El objetivo de esta tarea es que construyas una función para identificar la expresión regular

```
C[VA][ST]E.ISF[LIVM]T[SGC]EA[SCN][DE][KRQ]C
```

,

que corresponde al dominio de proteínas Hap2 que reconocen la caja CCAAT. Por otro lado, esta función será compatible con cualquier secuencia proveniente de UniProtKB para identificar expresiones regulares, por ejemplo motivos o dominios, aunque cabe mencionar que por alguna razón los nombres de algunos algunos organismos vienen delimitados por corchetes (**[]**), entonces antes de meter el archivo fasta a la función hay que eliminarlos. Los de esta tarea no presentan este problema.

Entrega: **Debes entregar una bitácora electrónica detallando paso a paso cada uno de los incisos, se evaluará la sintaxis y ejecución de los comandos, la salida de los bucles y la interpretación del bucle (describiendo porqué lo hiciste de esa forma). Y antes de entregar la tarea comprueba que todos los comandos corren correctamente. También debes enviar los gráficos que generaste.**

Enviar al correo: **pcr2.1@hotmail.com**

Plazo para entregar la tarea: **hasta las 23:59 del martes 28, pasado ese tiempo ya no lo calificaré.**

1. Crear una función que de argumentos reciba un archivo fasta (archivo.fasta) y un patrón (Dominio, Motivo).
2. Y que la función devuelva un **diccionario** que contenga la siguiente información:
   1. **Como clave**: identificador UniProt de la poteína.
   2. **Como objeto o valor** (en una lista) con los siguientes elementos:
   - Un contador
   - Nombre de la proteína
   - Organismo
   - Longitud de la proteína
   - Dominio identificado
   - Longitud del dominio

- Positión inicial del dominio
- Posición final del dominio

3. Para crear la función usa el archivo .fasta **UniProt_sequences.fasta** con el patrón:

```
C[VA][ST]E.ISF[LIVM]T[SGC]EA[SCN][DE][KRQ]C
```

4. A partir del diccionario determina la frecuencia de los géneros identificados y representa los más abundantes en un gráfico de barras y guárdalo.
   Recomendación: Para saber la frecuencia puedes usar la función **Counter**.
5. Una vez que estés satisfecho con el resultado de tu función ahora valídala usando el archivo: **UniProt_sequences2.fasta**, con el nuevo patrón:

```
'G[LIVMFY].{1,3}[AGCY][NASMQG].C[FYWC][LIVMFCA][NSTAD][SACV].[LIVMSF][QF]'
```

6. Finalmente, como en el inciso 4, usando la función **Counter**, determina cuántos dominios únicos se identificaron a partir de cada archivo fasta, de cada grupo representa los dominios más abundantes en un gráfico de barras y guárdalo.
   Recomendación: en este caso de los valores del diccionario usarás los dominios, algo parecido al inciso 4.

Los archivos de la tarea los descargas a partir de:

https://raw.githubusercontent.com/Bioinformatica2020/tareas/master/UniProt_sequences.fasta

https://raw.githubusercontent.com/Bioinformatica2020/tareas/master/UniProt_sequences2.fasta

Recomendación:

Los nombres de las proteínas los extraes con el siguiente patrón (aunque falta una edición final):

```
"#[ A-Z0-9,()':/\\\.[\]+a-z_-]{0,1000} OS="
```

Al resultado le eliminas `#` y `\sOS=` (`\s`: es un espacio en blanco) para quedarte finalmente con el nombre de la proteína, revisa los identificadores para que observes lo que sucede.

Recomendación: Haz uso de expresiones regulares, operadores, estructuras de datos y condicionales para resolver los incisos, revisa el contenido de las bitácoras ya que la información puede guiarte. Recuerda agregar encabezados en Markdown y comentarios en el código de lo que estás haciendo y explicando porqué lo estás haciendo de esa forma. Recuerda, los acentos, espacios y la letra ñ no están permitidos en variables, solo en los comentarios.

Referencias

1. Dorn A, Bollekens J, Staub A, Benoist C, Mathis D. A multiplicity of CCAAT box-binding proteins. *Cell*. 1987;50(6):863-872. doi:10.1016/0092-8674(87)90513-7
2. Li X yan, Mantovani R, Hooft Van Huijsduijnen R, Andre I, Benoist C, Mathis D. Evolutionary variation of the CCAAT-binding transcription factor NF-Y. *Nucleic Acids Res*. 1992;20(5):1087-1091. doi:10.1093/nar/20.5.1087
3. McNabb DS, Pinto I. Assembly of the Hap2p/Hap3p/Hap4p/Hap5p-DNA complex in Saccharomyces cerevisiae. *Eukaryot Cell*. 2005;4(11):1829-1839. doi:10.1128/EC.4.11.1829-1839.2005
4. Mao Y, Chen C. The Hap Complex in Yeasts: Structure, Assembly Mode, and Gene Regulation. *Front Microbiol*. 2019;10. doi:10.3389/fmicb.2019.01645