TAREA1

HERRAMIENTAS BIOINFORMÁTICAS PARA LAS CIENCIAS BIOLÓGICAS:
INTRODUCCIÓN A PYTHON

Introducción

Las cajas son secuencias de DNA reguladoras localizadas corriente arriba del inicio de la transcripción [1]. En *S. cerevisiae*, la caja **CCAAT** se ha encontrado en genes que codifican citocromos, y en eucariotas superiores se ha encontrado en genes que regulan el ciclo celular [2]. En hongos, el complejo Hap (Hap2, Hap3, y Hap4) es un activador transcripcional que se une a estas secuencia reguladoras, y está implicado en la expresión de genes asociados a la respuesta a estrés, homeostasis de hierro y virulencia [3]. Por otro lado, la caja **AGGGG**, también llamada elemento de respuesta a estrés (*STRE: stress response element*), se ha estudiado en algunos hongos, incluyendo al patógeno *Candida albicans*, y esta caja es responsable de la regulación de genes implicados en la respuesta a varios tipos de estrés [4] [5] [6].

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.**
Enviar al correo: **pcr2.1@hotmail.com**
Plazo para entregar la tarea: **hasta las 23:59 del domingo 19, pasado ese tiempo ya no lo calificaré.**

1. Descarga las secuencias promotoras del organismo asignado.
   Link para descargar las secuencias de *S. cerevisiae*: descarga
   Link para descargar las secuencias de *Aspergillus fumigatus*: descarga
   Recomendación: deposita el archivo en la carpeta de descargas que está dentro de Bio2020, al descargar el archivo cuida que no se pierda la extensión .fasta
2. Abre, edita y guarda en una variable iterable (lista) todas las secuencias promotoras del organismo.
   Recomendación: identifica patrones en el encabezado de las secuencias, edítalas de tal modo que tengan la estructura que se muestra en seguida para:

   S. cerevisiae

```
'851230|SEO1#AAATCGGAATTGGAGGTATCGGATCTTGTTGAATATCCACCAATGTCTTACCCCTGTATTTTAACAAGAGTTTACGCTGTTATATGGTTAAAGGTGTGGACGCCTTG
AAGGTTTACCTTACCGAATGACACCTTTACAATAGTCAGATCACGTTCTGTGGCGTTATCCAAAGTTAGCGCAGTTTTCCGATGGTCCAATGTAATCATTAGAAATAGTAAAAACTGTGTA
ATGGTAAAGATTGTGTCACTGGAAAAAAACTGCTACAAATAATAAATAAATAAAAAAATACGAAAGCACAGTACTACGGGTGCCTCCACAAATAGATAAGAAACCAAGCGGAGACATGCGT
TTAGATGAGGATATAAATTATTTATACAACCAGACTATATAAAAGAGCATCTAGTTTACCTGTTATGATGAATGGACATTCGCTACATATCTTACTCTCTATTTGTTAAAAAAAATTACAA
AGAGAACTACTGCATATATAAATAACATAC',
```

   A. fumigatus

```
'AFUA_1G00210#TCAATGTAAGGGTCATGGATTCAGCTTGCCTCATTACCGAATCACTTTGAGACTATGGAGCATTAACCAATCCGCACATAGGGTTTGAAGAAGGCAACAAACGCCA
AATAGCGCCAAGGGTTTTTTAATAGGCCCATGCCAGAGCGAGACTCTGGCCGTCGGCGAACCCAAGAAAAGAAATGAGAACCAGCCTAGGACTTATTATTTAGAAAATAACTCACAATGTG
CTCTCATAGCACCATCTATTTCTGGGGGGTATACGACTAAATCAATAGGTGGCCAAGTAACGGCTTGGTCTAACCATTATAGGCGAATCCCATTGGGTTGTTCCAAATCGAGAGGCGCACG
GAGCTTGTTAGCACAGCCCAATGGGTTGTTCCTGATTGAGGAAAGATCGACGCAATTCAATCCAATTGGACTTCTGCAAGATCTGAATGCTAGCAAAATGCAATTGACACTCAGCCTGGCA
GCGGTCCAAGGGTAGAATTTGCCATACACTC',
```

3. Usa la variable iterable en un bucle **for** para localizar la caja **CCAAT** en las secuencias promotoras.
4. En el cuerpo del bucle **for** imprime un contador, el identificador de la secuencia, las cajas y la cantidad de cajas encontradas por secuencia.
5. Guarda en una lista solo los identificadores que salen del bucle **for**.

   Recomendación: puedes guardar la lista en una variable con el nombre de la caja.
6. Realiza el mismo procedimiento (inciso 2, 3 y 4) pero ahora con la caja **AGGGG**.
7. Usa los identificadores que están en las listas para construir dos conjuntos, uno por caja, y responde:
   - Qué genes y cuántos tienen ambas cajas dentro de sus promotores.
   - Haz un diagrama de Venn para visualizar la intersección entre los dos conjuntos de genes.

   Recomendación: salva el diagrama de Venn en formato .png y lo envías junto con la bitácora.
8. Por alguna razón el algoritmo de RSAT (*Regulatory Sequence Analysis Tools*: http://rsat-tagc.univ-mrs.fr/rsat/) para extraer secuencias promotoras devuelve identificadores sin secuencias, identifica cuántos identificadores no tuvieron secuencias.
9. **Recomendación:** esto se resuelve con un buble **for** usando la variable iterable generada en el inciso 1.

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. Adcock IM, Caramori G. Transcription factors. In: *Asthma and COPD*. Elsevier Ltd; 2009:373-380. doi:10.1016/B978-0-12-374001 4.00031-6
2. Wasner M, Haugwitz U, Reinhard W, et al. Three CCAAT-boxes and a single cell cycle genes homology region (CHR) are the major regulating sites for transcription from the human cyclin B2 promoter. Gene. 2003;312:225-237. doi:10.1016/s0378-1119(03)00618-8
3. 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
4. eterbauer CK, Litscher D, Kubicek CP. The Trichoderma atroviride seb1 (stress response element binding) gene encodes an AGGGG-binding protein which is involved in the response to high osmolarity stress. Mol Genet Genomics. 2002;268(2):223-231. doi:10.1007/s00438-002-0732-z
5. struch F. Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. FEMS Microbiol Rev. 2000;24(4):469-486. doi:10.1111/j.1574-6976.2000.tb00551.x
6. Nicholls S, Straffon M, Enjalbert B, et al. Msn2- and Msn4-like transcription factors play no obvious roles in the stress responses of the fungal pathogen Candida albicans. Eukaryot Cell. 2004;3(5):1111-1123. doi:10.1128/EC.3.5.1111-1123.2004