Universidad de los Andes
Facultad de Ingeniería

Facultad de Ciencias

Maestría en Biología Computacional
BCOM4006 – Algoritmos en Biología Computacional
Semestre: 2024-20

Tarea 1 – Bases de datos de ADN

Objetivos

1. Practicar algunos análisis comunmente realizados en genómica comparativa
2. Conocer algunas de las bases de datos de ADN
3. Practicar programación como herramienta para entender el proceso de traducción de ARN en proteínas.

Instrucciones

El gen GBSSI es uno de los genes más importantes a estudiar en relación con el contenido de almidón en varias especies de plantas. En esta tarea se busca realizar un análisis de variabilidad de este gen en diferentes especies utilizando la información disponible en diferentes bases de datos, tanto de genomas completos como de secuenciación dirigida. Para realizar esta tarea, siga los siguientes pasos:

1. [30%] Utilizar la función BLAST disponible en diferentes bases de datos de ADN para buscar alelos secuenciados del gen GBSSI. Consultar por lo menos las siguientes bases de datos:
a. NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
b. Uniprot (http://www.uniprot.org/)
c. Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html).

Se adjunta una copia de la secuencia de nucleótidos del gen en formato fasta para facilitar el proceso de búsqueda por medio de consultas blast en estas o en otras bases de datos relevantes. Para el caso de NCBI, describir qué recibe y qué retorna cada uno de los 4 tipos de BLAST disponibles en la página principal. Para las demás bases de datos, describir qué tipo de BLAST fue ejecutado. Se debe generar un archivo fasta con las secuencias de nucleótidos encontradas en la búsqueda. Asegurese que las secuencias descargadas corresponden solamente a ARN maduro, es decir, que no contengan intrones. La información completa sobre el formato FASTA se puede consultar en este enlace:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp

2. [20%] Desarrollar un programa que reciba el archivo generado en el primer punto y retorne un nuevo archivo fasta que filtre el archivo original de modo que no contenga secuencias repetidas.

3 [30%] Desarrollar un programa que reciba un archivo fasta con secuencias de nucleótidos como el generado en los puntos anteriores y que genere un archivo fasta con secuencias de aminoacidos que representen la traducción a proteina de las secuencias presentes en el archivo de entrada. Para cada secuencia es importante identificar el codón de inicio para comenzar la traducción y, una vez avance la traducción, identificar los codones de parada para terminarla. Utilizar este programa para generar las secuencias de aminoácidos correspondientes a las secuencias de nucleótidos extraidas en el segundo punto. Discutir si utilizando secuencias sin repetidos como entrada se puede garantizar que no haya secuencias repetidas en la salida. Para complementar las secuencias de aminoácidos construidas en este punto, encontrar secuencias de aminoácidos utilizando el tipo adecuado de BLAST en las bases de datos mencionadas arriba.

4. [20%] Realizar un alineamiento múltiple de secuencias con las secuencias de nucleótidos y otro con la secuencia de aminoácidos recolectadas en los puntos anteriores. Pueden utilizar el algoritmo MUSCLE disponible en el sitio web de EBI (http://www.ebi.ac.uk/Tools/msa/muscle/) o cualquier otra alternativa que les permita generar un alineamiento de la mejor calidad posible. Describir cual de los dos alineamientos parece tener mejor calidad y qué filtros se podrían realizar a las secuencias iniciales para mejorar la calidad del alineamiento. Describir el formato ClustalW para guardar alineamientos de secuencias.

5. (Bono [10%]) Investigar herramientas que permitan construir dendogramas con el algoritmo UPGMA y con el algoritmo Neighbor Joining a partir de cada uno de los alineamientos construidos en el punto anterior. Ejecutar estas herramientas y mostrar los dendogramas correspondientes. Incluir la referencia correcta a la herramienta que se utilizó para ejecutar estos análisis. Describir las diferencias más relevantes que se vean entre los diferentes dendogramas.

Entrega

Se debe entregar un informe breve describiendo las acciones realizadas para seguir cada uno de los pasos y los resultados encontrados. Se debe adjuntar también el código fuente desarrollado. Generar un solo archivo .zip con todos los entregables del trabajo. Para evitar problemas de compatibilidad, el informe se debe entregar en formato PDF.