UNER

Facultad de Ingeniería
Lic. en Bioinformática

Análisis y Alineamiento de Secuencias

Trabajo Práctico Integrador

2022

En el presente Trabajo Práctico Integrador se abordarán principalmente las temáticas de ensamblado y anotación de secuencias, retomando también conceptos vistos en la primera parte del cursado de la asignatura.

Objetivos

- Aplicar los conceptos teóricos y prácticos vistos en la materia para el ensamblado y anotación de una parte de un genoma.
- Profundizar el manejo del *software* específico para la temática.
- Profundizar la programación de *scripts* en el lenguaje Perl.

Introducción

Hoy en día, con los crecientes avances en la genética y los aportes bioinformáticos, se puede realizar el ensamblado y anotación de genomas de distintas especies. En particular, la secuenciación automática de ADN permite obtener tramos de secuencias que raramente sobrepasan los 800 pb. Estas secuencias tienen que ser posteriormente ensambladas para formar *contigs* (conjunto de secuencias con regiones superpuestas), los cuales son nuevamente ensamblados para formar *contigs* mayores (*scaffolds* o *supercontigs*). Este ensamblado requiere de regiones de superposición de las secuencias y de programas específicos que identifiquen dichas regiones y alineen en forma secuencial los distintos tramos de secuencias obtenidos.

Ustedes forman parte de un grupo de investigación involucrado con el proyecto de resecuenciación del genoma de *Brucella suis*. Los biotecnólogos del grupo han obtenido electroferogramas procedentes de clones individuales de una biblioteca construida con ADN fraccionado al azar de la bacteria en cuestión. A cada equipo de investigación del área bioinformática se le brinda un Set de aproximadamente 150-300 electroferogramas, y los investigadores principales necesitan que los mismos sean ensamblados, y en todos los casos, deben finalizar con la construcción de un *contig* cuya secuencia consenso posee entre 20.000-25.000 pb.

Actividades

1- Para interiorizarse en el organismo que están estudiando, se les pide realizar un pequeño informe bibliográfico acerca de las bacterias del género *Brucella*, en particular de *B. suis*, teniendo en cuenta información vinculada con la composición de su genoma y ensamblados existentes, y también con su impacto en la salud humana y animal, patogenia, incidencia económica, y demás particularidades que consideren relevante. No pierdan de vista que al conocerse la secuencia completa de *B. suis* se obtiene información sobre el estilo de vida, patogenia y evolución de este patógeno.

A partir de los archivos de secuencia provistos:

2- Procesar las secuencias para eliminar las regiones de baja calidad y vector. Ensamblar las secuencias y obtener como resultado un solo contig. En caso que esto sea imposible, proseguir con el problema con todos los contigs mayores a 10 kb.

3- Obtener la secuencia del contig en formato fasta. Reportar el tamaño y el %G+C. ¿Por qué es importante el contenido de G+C en un organismo procariota?

4- Con la secuencia obtenida, realizar una predicción de genes. Se recomienda utilizar una combinación de los resultados.

Por ejemplo, para realizar una predicción con un programa como Glimmer, utilizar alguno de los genomas de género *Brucella* para entrenar el HMM.

En caso de elegir un programa, o más de uno, justifiquen por qué. ¿De qué forma consideran que es más confiable y conveniente realizar una predicción de genes?

5- Obtener dos archivos en formato fasta con las predicciones de los genes. Uno de ellos debe contener las secuencias de los genes, y el otro las secuencias traducidas a proteína (nombrarlos, enumerándolos como Set_01, Set_02, etc., para cada uno de los genes/proteínas predichos en el Set 1, y así respectivamente de acuerdo al que les tocó).

6- Comparar las proteínas predichas contra bases de datos para realizar su anotación (o bien, refinarla si utilizaron un *pipeline* integrador en el paso 4). Pueden utilizar el script de BLAST remoto en Perl o pueden usar la interfaz web para ver los gráficos, pero el resultado en formato de texto debe ser generado y entregado.

7- Realizar un script en Perl que genere una tabla (texto separado por tabulaciones "\t") que contenga los primeros 10 hits para cada comparación de la actividad anterior.

Cada archivo de salida debe nombrarse con el nombre de la proteína (paso 5), y debe tener el formato:

Hit Nro.	Descripción	% identidad	Inicio Match	Fin Match	% secuencia alineada

NOTA: Utilizar el módulo Bio:SearchIO de BioPerl, cuya documentación se encuentra en https://bioperl.org/howtos/SearchIO_HOWTO.html

Además consultar: https://bioperl.org/howtos/Beginners_HOWTO.html

Realizar una predicción de función para cada proteína cuando esta sea posible y reportarlo.

Si en el paso 4 utilizaron un *pipeline* integrador (predicción + anotación), ¿Qué similitudes y/o diferencias encuentran con los resultados obtenidos en este paso?

8- Seleccionar una proteína con función predicha y obtener las secuencias completas de los primeros 10 hits de PSI-BLAST (3 iteraciones) en los cuales se muestren similitudes con otras especies además del género *Brucella*.
Realizar un alineamiento múltiple entre todas estas secuencias utilizando ClustalX.

NOTA: además de que los resultados de BLAST muestren similitudes con otras especies, verificar que este alineamiento contenga una buena parte de la proteína, ya que sino el alineamiento múltiple global no tendría sentido.

Metodología de evaluación

Se debe entregar una carpeta con todos los archivos utilizados: el ensamblado, los archivos .fasta generados, las tablas y otros resultados solicitados. Se pueden utilizar los scripts en Perl generados durante el cursado. Adicionalmente, debe entregarse un reporte de las actividades realizadas.

Se disponen de 3 semanas para la realización del trabajo, el cual debe ser subido al campus en un archivo comprimido que contenga tanto al reporte como a todos los archivos solicitados.

La fecha de entrega límite es **el jueves 27 de octubre a las 11 hs**. Ese mismo día se deberá exponer, contando al resto de la clase qué hicieron y los resultados que obtuvieron.

Del informe escrito presentado se evaluarán los resultados obtenidos y los criterios con los que se llevó a cabo. Es decir, se pide que los resultados obtenidos sean coherentes con lo que se ha trabajado en clase, las respuestas acordes a la temática y completitud de las mismas, autonomía en el desarrollo de *scripts* en Perl, justificación pertinente de los programas elegidos, funcionalidad de los archivos generados, etc. Se evaluará la participación en la puesta en común y el manejo adecuado del contenido y vocabulario.