Análisis y alineamiento de secuencias

Año: 2022

Trabajo práctico integrador

“Brucella suis”

Salim Taleb, Nasim A; Mansilla, Leandro

Docentes: Rausch, Atilio; Tossolini, Ileana

Carrera: Lic. en Bioinformática

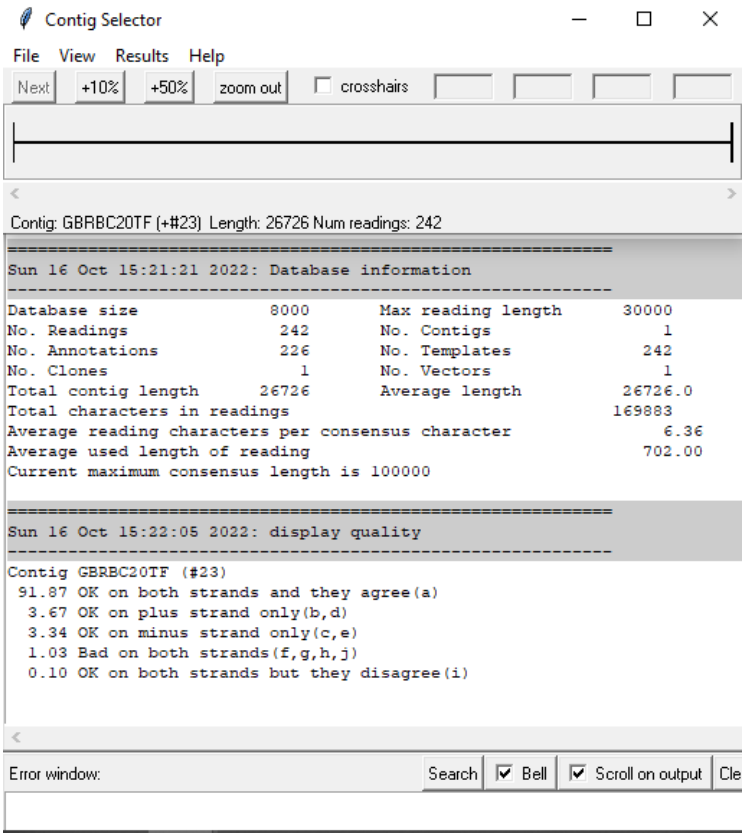
Desarrollo

1) Brucella Suis es una bacteria gram negativa del género Brucella específica de los cerdos, causa problemas reproductivos en los mismos y en humanos puede ocasionar una respuesta inmune como fiebre. El contagio humano usualmente ocurre por un contacto accidental con animales infectados o con la ingesta de comida contaminada. El organismo es considerado un potencial agente bioterrorista y fue uno de los primeros organismos patógenos en ser usado como arma por EEUU.

Una vez que el organismo entra al cuerpo, se vuelve intracelular y entra en la sangre y las regiones linfáticas, multiplicándose dentro de los fagocitos y eventualmente causando bacteriemia.

Existen unos 82 alineamientos al dia de la fecha en el NCBI de este organismo, también se cuentan con otros datos de interés como por ejemplo la mediana de largo total de secuencia de 3,31581 Mb, la mediana de cantidad de proteínas de 2967 y la mediana de %GC de 57,2%. Fuente: https://www.ncbi.nlm.nih.gov/genome/?term=Brucella%20suis

2) Para el procesamiento de cada una de las secuencias provistas en el dataset se utilizó pregap4. Posteriormente se ensamblaron las secuencias mediante gap4 dando los siguientes resultados:



Obteniendo como resultado del ensamblado y de unir los contigs, un sólo contig de un largo de 26726 pb, contando pads y gaps. Se creó un archivo “consenso.fasta” con la secuencia.

3) Luego de extraer la secuencia consenso de gap4 en formato FASTA se utilizan dos herramientas online para obtener el tamaño total de la secuencia y el porcentaje de GC:

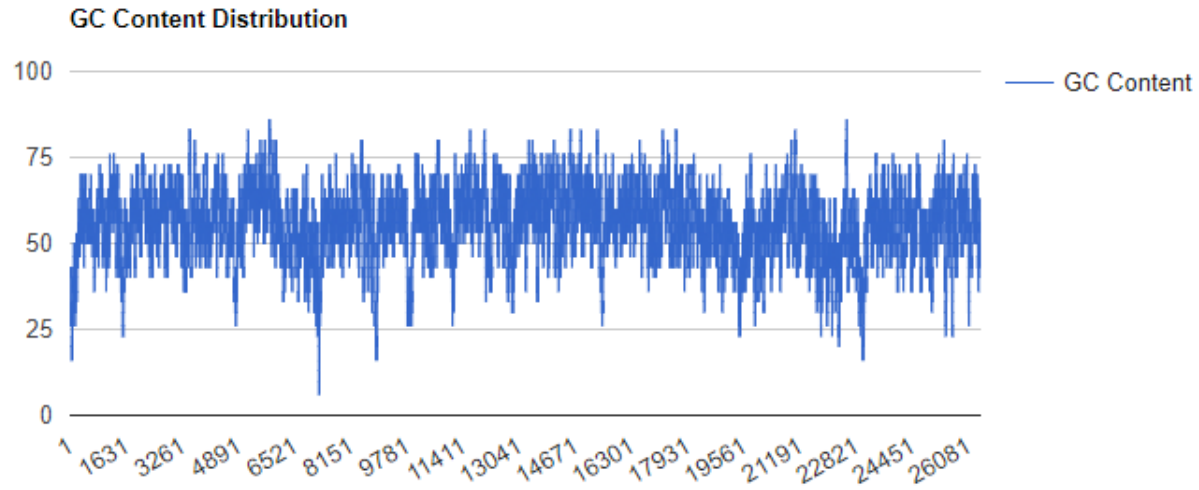
Results

GC content: 55.95%

Base	Count	Percentage
A	6096	23.06
C	7658	28.97
G	7130	26.98
T/U	5546	20.98

https://jamiemcgowan.ie/bioinf/gc.html

GC Content Calculator



Summary: Full Length(26430bp) | A(23% 6096) | T(23% 5546) | G(26% 7130) | C(28% 7658)

https://www.biologicscorp.com/tools/GCContent/

Concluyendo que la secuencia consenso tiene un largo de 26430 pb y un porcentaje de GC de 55,95%.

Este porcentaje de GC resulta importante en el caso de los organismos procariotas ya que es un carácter ampliamente utilizado en las descripciones taxonómicas de estos.

4) Para la predicción de genes de la secuencia consenso obtenido se optarán por 3 alternativas, ya que siempre lo mejor es tener múltiples resultados de una predicción y anotación para poder compararlas y no cometer errores u omisiones, se usará Prokka, Bakta y finalmente FgenesB para la predicción y luego una búsqueda en BLAST para la anotación ya que FgenesB sólo realiza una predicción.

FgenesB:

Como parámetro se utilizó la tabla de código genético 11, “The Bacterial, Archaeal and Plant Plastid Code”, al seleccionar el organismo más cercano se observó que existen dos opciones para Brucella suis, chromosome I y chromosome II, por lo tanto se realizó una corrida con cada una y se consideraron predicciones diferentes, los resultados obtenidos fueron guardados en “reschromosome1.txt” y “reschromosome2.txt”, analizandolos resultan ser ligeramente diferentes y también se detectaron 25 genes utilizando cromosoma 1 y 28 con el cromosoma 2.

Prokka:

Para el caso de prokka se corre simplemente el comando “prokka consenso.fasta --evalue 1-e05” y se obtienen los archivos de resultados, en el archivo “PROKKA_10232022.txt” se escribe la cantidad de CDS encontrados, en este caso 24.

Bakta:

Para el caso de Bakta se optó por Bakta web (https://bakta.computational.bio/), una vez finalizado el proceso se muestra una página con los resultados, similar a BLAST del NCBI, en la página se muestra que se encontraron 38 CDS.

5) FgenesB:

La salida del programa ya cuenta con los genes en formato fasta en proteínas, para obtener la secuencia de nucleótidos se utilizó “extractseq” del paquete Emboss con las posiciones obtenidas, también se utilizó “revseq” de Emboss para obtener la reversa complementaria y extraer las secuencias en la hebra negativa.

Prokka:

Para Prokka no es necesario ningún tratamiento ya que el archivo “PROKKA_10232022.ffn” contiene los distintos genes encontrados en secuencia de nucleótidos y “PROKKA_10232022.faa” en secuencia de proteínas.

Bakta:

De manera similar a Prokka, Bakta Web permite bajar los archivos con extensión “ffn” y “faa” que contienen las secuencias de ADN y aminoácidos respectivamente.

6) Para cada set de resultados, los 2 de fgenesB, Prokka y Bakta, se ejecuta una búsqueda en BLAST web utilizando BLASTP y la base de datos nr y se adjuntan los resultados en cada carpeta correspondiente, posteriormente se cruzan y revisan manualmente los resultados de BLASTP y los de cada predicción de fgenesB y se anotan los genes, algo similar se hace con las salidas de Prokka y Bakta sólo que al ser pipelines integradores algunos de los genes ya están anotados.

7) Para poder procesar los resultados de BLASTP mediante el script y la librería Bio::SearchIO se tuvo que descargar nuevamente los resultados de la página de BLAST del NCBI pero está vez desde la interfaz del antiguo formato, ya que los obtenidos en la actividad 6 para la anotación no se leían correctamente.

Consideraciones de uso del script:

- Se coloca el nombre del archivo de entrada por consola al ejecutar el script.
- Sólo se procesa el mejor HSP por hit.
- Si una proteína no tiene hits no se crea un archivo de salida y se omite en el procesamiento.
- Se reemplazan "/" en los nombres de las proteínas por "-" ya que genera conflictos al abrir los archivos.
- Se consideró a %secuencia alineada como la fracción de la secuencia que estaba "conservada", es decir, el porcentaje de bases que tuvieron un score positivo en el alineamiento.

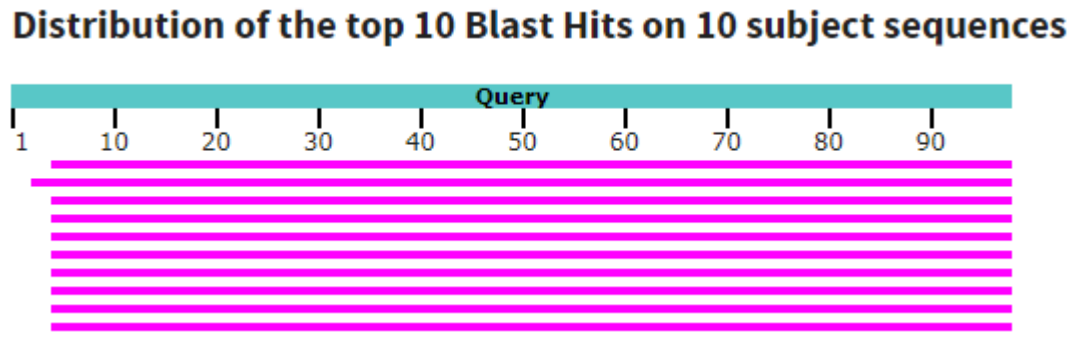
Comparación de resultados

De manera general los archivos de salida obtenidos mediante pipelines integradores requirieron poca refinación en cuanto a la anotación, a diferencia de los obtenidos con fgenesb que requirieron anotarlos por completo de manera manual y por lo tanto un mayor esfuerzo. Bakta resultó más precisa que Prokka en cuanto a la anotación, es decir, que requirió una menor refinación, pero detectó mayor cantidad de proteínas hipotéticas. La predicción de genes con el cromosoma I de fgenesb resultó en una menor cantidad de proteínas hipotéticas, 2, que la predicción con el cromosoma II, 6, lo que podría indicar que el fragmento pertenece al cromosoma I ya que la predicción de genes con él obtuvo mayor proporción de proteínas ya identificadas.

Consenso final

Para llegar a la anotación final se revisan cada una de las proteínas obtenidas por cada método haciendo alineamientos múltiples mediante ClustalO, un predecesor de ClustalW, (https://www.ebi.ac.uk/Tools/msa/clustalo/) para tener un mejor panorama, el criterio para anotar de manera definitiva una proteína será de obtener una secuencia consenso que tenga la mayor cantidad posible de identidad con las demás secuencias. Se omitirán proteínas que no se repitan en al menos dos anotaciones de la anotación final, por ejemplo., Bakta arrojó una gran cantidad de proteínas hipotéticas pero no se encuentran la mayoría en las demás anotaciones, tampoco se considerará a las anotadas únicamente por ambos fgenesb ya que son muy similares porque provienen de la misma fuente de anotación. Se guardan los resultados de los alineamientos y en cada uno se aclara que criterio se tomó para la anotación final, de ser necesario.

8) La proteína seleccionada es la última de la anotación consenso “23-glutamato descarboxilasa piridoxaldependiente”. El resultado gráfico de las 10 mejores secuencias después de la tercera iteración es el siguiente:



Para alinear estas 11 secuencias se utilizaron múltiples herramientas diferentes, la previamente usada ClustalO, ClustalW (https://www.genome.jp/tools-bin/clustalw) y T-cofffee (https://www.ebi.ac.uk/Tools/msa/tcoffee/). Se guardaron todas las salidas en al carpeta “Act8”.