Taller6_AnotacionProcariota

July 17, 2025

1 Taller Anotacion con Prokka (Procariotas)

Estudiante: Francisco Salamanca

En este tutorial realizaremos una introducción al uso de PROKKA para la anotación de un genoma procariota. Para esto usaremos la implementación de Prokka en la plataforma Galaxy y en local.

Prokka (Rapid Prokaryotic Genome Annotation) es una herramienta bioinformática desarrollada para la anotación rápida, precisa y estandarizada de genomas procariotas. Fue diseñada por Torsten Seemann y es ampliamente utilizada en genómica microbiana para predecir genes y asignarles funciones con un alto grado de automatización. Prokka integra múltiples herramientas y bases de datos especializadas, lo que le permite generar anotaciones estructurales y funcionales de manera coherente y reproducible.

A partir de una secuencia genómica ensamblada en formato FASTA, Prokka realiza los siguientes procesos principales:

1. Predicción estructural de elementos genómicos, incluyendo:

- Genes codificantes (CDS)
- ARN ribosómicos (rRNA)
- ARN de transferencia (tRNA)
- Genes hipotéticos

2. Asignación funcional mediante la comparación contra bases de datos de proteínas de referencia, tales como:

- UniProt (Swiss-Prot/TrEMBL)
- RefSeq
- COG
- Base de datos específica por género si se activa la opción –usegenus

3. Generación de archivos estandarizados en múltiples formatos para su posterior análisis o presentación a bases de datos públicas, incluyendo:

- .gff (General Feature Format)
- .gbk (GenBank)
- .faa (secuencias proteicas)

- .ffn (secuencias nucleotídicas de genes)
- .sqn (para envío a GenBank, si se configura tbl2asn)

1.1 Desarrollo del taller

1.1.1 Parte 0: Instalación

Instalacion de prokka mediante Docker, para evitar instalar en local, y evitar errores de versiones.

[ ]:
```
!docker pull staphb/prokka
```

1.1.2 Parte 1: Correr Prokka con Docker

Descripción general del uso de Prokka:

1. Prepare la entrada: necesita un único archivo FASTA que contenga los contigs ensamblados de su genoma (por ejemplo, my_genome.fasta). La entrada debe estar en formato FASTA, que es el formato de salida de una metodología de ensamblaje denovo, en lugar de uno de referencia (.bam)
2. Ejecute el comando: El comando básico para ejecutar Prokka es:

   `prokka my_genome.fasta`
3. Personalice la salida: Puede utilizar indicadores para organizar sus resultados: * –outdir : Indica a Prokka dónde guardar los archivos de salida. Por ejemplo, –outdir prokka_results. * –prefix : establece un nombre específico para todos los archivos de salida. Por ejemplo, –prefix my_organism. * –kingdom : especifica el reino (Archaea, Bacteria o Viruses).

   Un comando más completo sería así:

   `prokka --outdir prokka_results --prefix my_organism --kingdom Bacteria my_genome.fasta`
4. Revisa los resultados: Prokka creará un nuevo directorio (por ejemplo, prokka_results) que contiene varios archivos de salida. Los más importantes son: * .gff: un archivo de formato estándar con todas las anotaciones, que se puede ver en un navegador genómico. * .gbk: un archivo de formato GenBank, también muy utilizado. * .faa: un archivo FASTA con las secuencias de proteínas previstas. * .txt: un archivo resumen con estadísticas sobre la anotación (por ejemplo, número de genes encontrados).

[ ]:
```
#Anotar genes, proteinas y otros. usando prokka

!docker run --rm -v $(pwd):/data staphb/prokka prokka /data/final.contigs.fa␣
 ↪--outdir /data/anotacion_prokka --prefix mycoplasma --kingdom Bacteria␣
 ↪--force

#Esto ejecuta Prokka dentro del contenedor, pero usando los archivos locales.

#--rm: elimina el contenedor después de que termina (no ocupa espacio).

#-v $(pwd):/data: monta el directorio actual como /data dentro del contenedor.
```

```
#staphb/prokka: nombre de la imagen Docker que tiene Prokka instalado.

#prokka ...: es el comando normal de Prokka, solo que los paths son relativos a␣
 ↪/data.
```

1.2 Preguntas Orientadoras

Para este tutorial deberá desarrollar un informe escrito que posteriormente deberá subir a través del classroom de la clase. Para este informe tenga en cuenta las siguientes preguntas orientadoras:

1. **¿Qué parámetros específicos de Prokka se ha configurado para reflejar las características del genoma de Mycobacterium?**

R/

El unico parametro especifico configurado al correr proka fue – kingdom Bacteria, especificando que el genoma a anotar proviene de el reino Bacteria, cabe destacar que en el archivo .err de proka se encuentran inconsistencias como:

- Missing protein IDs
- Inconsistent gene locations
- Suspect product names
- Short rRNA features
- Non-standard rRNA product names

2. **¿Cuántos genes codificantes (CDS) ha identificado Prokka?**

R/ Se encontraron un total de 1004 genes codificantes, esta informacion se ubica en el .log, al igual que en el .err.

3. **¿Qué información contienen los archivos .gff, .gbk, .faa y .ffn? Explique su utilidad para análisis posteriores.**

R/

- .gff (General Feature Format): Es un archivo separato por tabuladores, y describe todas las caracteristicas genomicas encontradas como tRNAs,rRNAs y tmRNAs, en conjunto con sus coordenadas y atributos.
- .gbk (GenBank): Es un archivo que combina la secuencia de nucleotidos de cada contig con cada anotacion predecida, incluye informacion del organismo, codigo geneticom y como se anoto.
- .faa (Fasta amino acid): Archivo que contiene la secuencia de aminoacidos de todas las secuencuas de proteinas predichas en un formato FASTA. Proteinas Predichas.
- .ffn (Fasta Nucleotide of features): Contiene secuencias de nucleotidos de todos los genes codificandes predichos y caracteristicas de RNA

4. **Por otra parte seleccione tres genes anotados por Prokka que usted considere biológicamente relevantes (por ejemplo, involucrados en patogenicidad, resistencia antimicrobiana o metabolismo lipídico). ¿Qué función tienen y cómo se relacionan con la fisiología de Mycobacterium?**

R/

1. Adhesin P1 (Locus Tag: JLEMDLKA_00006)

    - Función: La Adhesina P1 es una proteína de superficie principal en Mycoplasma pneumoniae (y otras especies de Mycoplasma patógenas). Su función principal es mediar la adhesión de la bacteria a las células epiteliales del huésped, particularmente en el tracto respiratorio. Esta adhesión es un paso crítico para la colonización y el establecimiento de la infección.

    - Relación con la fisiología de *Mycoplasma*: Para Mycoplasma, que carece de pared celular y es un parásito obligado, la adhesión a las células del huésped es fundamental para su supervivencia y patogénesis. La Adhesina P1 permite que la bacteria se ancle a las superficies celulares, resista los mecanismos de eliminación del huésped y obtenga nutrientes de su entorno, lo que la convierte en un factor de virulencia clave.

2. Phosphatidylglycerol–prolipoprotein diacylglyceryl transferase (Locus Tag: JLEMDLKA_00390)

    - Función: Esta enzima está implicada en la biosíntesis de lipoproteínas, catalizando la transferencia de diacilglicerol a prolipoproteínas. Las lipoproteínas son componentes importantes de la membrana celular de Mycoplasma.

    - Relación con la fisiología de *Mycoplasma*: Las lipoproteínas son abundantes en la superficie de Mycoplasma y desempeñan funciones cruciales en la interacción con el huésped, la adquisición de nutrientes y la modulación de la respuesta inmune. La actividad de esta enzima es esencial para la biogénesis de la membrana de Mycoplasma y, por lo tanto, para su viabilidad y capacidad de interactuar con su entorno y huésped.

3. Oligopeptide transport ATP-binding protein OppD (Locus Tag: JLEMDLKA_00379)

    - Función: OppD es un componente de un sistema de transporte ABC (ATP-binding cassette) conocido como el sistema de permeasa de oligopéptidos (Opp). Este sistema es responsable de la captación de pequeños péptidos (oligopéptidos) del entorno.

    - Relación con la fisiología de *Mycoplasma*: Las especies de Mycoplasma tienen capacidades biosintéticas limitadas y son auxótrofas para muchos nutrientes esenciales, incluidos los aminoácidos. Dependen en gran medida de la adquisición de estos nutrientes de su huésped. El sistema Opp, con OppD como un componente clave de unión a ATP, es vital para obtener péptidos del huésped, que luego pueden ser descompuestos en aminoácidos para la síntesis de proteínas y la generación de energía. Este sistema es crucial para la supervivencia y el crecimiento de Mycoplasma dentro del huésped.