Análisis del Genoma de *Escherichia coli*

Nombre: Hely Salgado (heladia@ccg.unam.com)

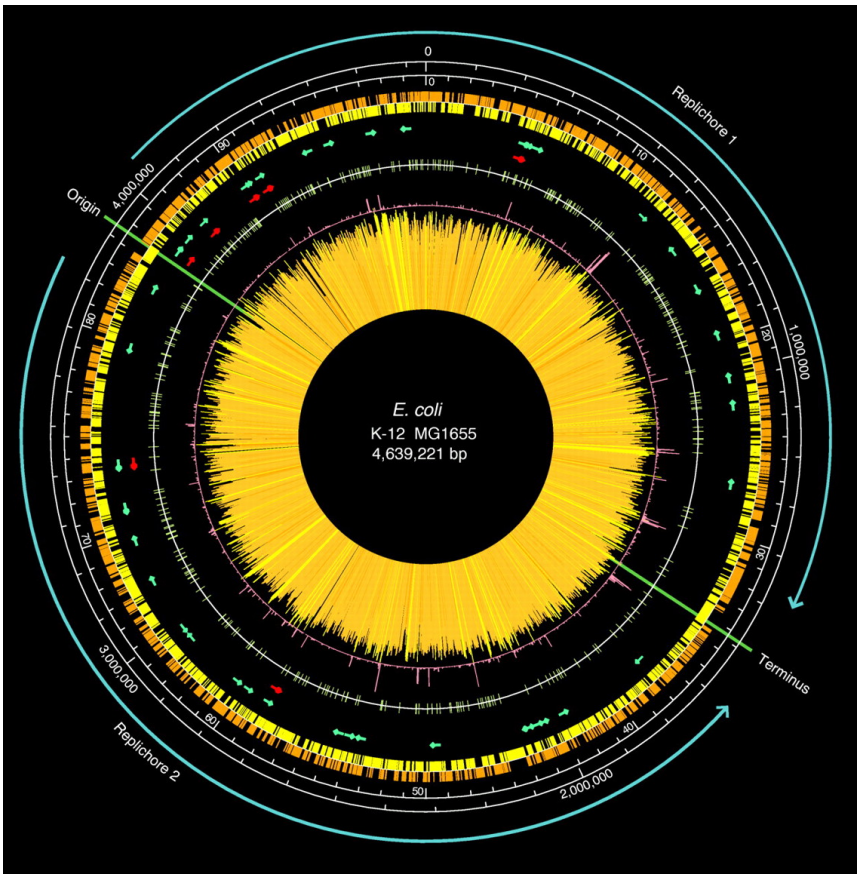
Fecha: 22 de sept 2023

Introducción

E. coli es la bacteria anaerobia comensal más abundante de la microbiota del tracto gastrointestinal, en donde junto con otros microorganismos es esencial para el funcionamiento correcto del proceso digestivo.

Escherichia coli también es un organismo modelo utilizado frecuentemente en el laboratorio por su velocidad de crecimiento, por sus pocos requerimientos nutricionales y por su amplia bibliografía. Además, E. coli es usada en experimentos de genética y biología molecular debido a que la estructura de su genoma es altamente flexible [1].

Debido a su extraordinaria posición como modelo preferido en genética bioquímica, biología molecular y biotecnología, E. coli K-12 fue el primer organismo sugerido como candidato para la secuenciación del genoma completo [2]. Su secuencia se completó y reportó en 1997.



Su genoma completo y sus anotaciones pueden ser descargadas desde NCBI.

Este análisis consiste en explorar y conocer un poco más sobre el genoma de *E. coli*, a través de una serie de preguntas planteadas que abordaremos usando comandos de unix para responderlas.

Al final daremos nuestra opinión sobre los hallazgos.

Metodología

A. Servidor y software

> Servidor: tepeu.lcg.unam.mx
>
> Usuario: compu2
>
> Software: sistema operativo unix

B. Datos de Entrada

Los datos de entrada han sido copiados desde /home/compu2/WelcomeBioinfo/datos/practica4, y éstos fueron descargados desde NCBI.

```
|-- data
|   |-- coli_genomic.fna
|   |-- coli.gff
|   |-- coli_protein.fna
|   |-- directorio.txt
|   `-- flagella_genes.txt
```

metadatos de la carpeta de datos

> Versión/Identificador del genoma: NC_000913.3
>
> Fecha de descarga:

Archivo	Descripción	Tipo
coli_genomic.fna	Secuencia de nucleotidos de E. coli	Formato FastA
coli.gff.	Anotación del genoma de E. coli	Formato gff
coli_protein.faa	Secuencia de aminoacidos de las proteinas de E. coli	formato FastA
flagella_genes.txt	Genes con función relacionada al flagello en E. coli	lista
directorio.txt.	Archivo con nombres de personas	lista

Formato de los archivos

- `coli_genomic.fna` : formato FastA

```
>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTG
GTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCA
AGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCA
```

Formato:

> a. La primera línea es información de la secuencia. Primero viene el identificador del

> genoma.
>
> b. Después vienen varias líneas con la secuencia de nuclótidos del genoma completo.

- `coli.gff` : anotación de features en el genoma

El contenido del archivo es

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM584v2
#!genome-build-accession NCBI_Assembly:GCF_000005845.2
##sequence-region NC_000913.3 1 4641652
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=51114

NC_000913.3     RefSeq  region  1       4641652 .       +       .       ID=
NC_000913.3     RefSeq  gene    190     255     .       +       .       ID=
NC_000913.3     RefSeq  CDS     190     255     .       +       0       ID=
NC_000913.3     RefSeq  gene    337     2799    .       +       .       ID=
NC_000913.3     RefSeq  CDS     337     2799    .       +       0       ID=
```

Formato:

> a. Es un formato gff tabular, es decir cada dato es separado por tabulador.
>
> b. Cada renglón en el formato gff es una elemento genético anotado en el genoma, que se le denomina `feature` , éstos features pueden ser genes, secuencias de inserción, promotores, sitios de regulación, todo aquello que este codificado en el DNA y ocupe una región en el genoma de E. coli.
>
> c. Los atributos de cada columna par cada elemento genético son
>
> ```
> 1. seqname. Nombre del cromosoma
> 2. source. Nombre del programa que generó ese elemento
> 3. feature. Tipo de elemento
> 4. start. Posición de inicio
> 5. end. Posición de final
> 6. score. Un valor de punto flotante
> 7. strand. La cadena (+ , - )
> 8. frame. Marco de lectura
> 9.  attribute. Pares tag-value, separados por coma, que proveen informa
> ```

Resultados

1. ¿De qué tamaño es el genoma de *Escherichia coli*?

Según la referencia Blattner et al. se secuenció el genoma completo de la bacteria *E. coli*. Para determinar cual es el tamaño del genoma haremos lo siguiente:

Archivo(s): data/coli_genomic.fna

Algoritmo:

> - Usar el archivo data/coli_genomic.fna que contiene la secuencia del genoma completo.
> - Contar los caracteres del archivo (Lo ideal seria tomar solo la secuencia)
> - Al total hay que restar el num de lineas del archivo (porque se cuenta el salto de linea), menos los caracteres del primer renglón que no es secuencia.

Solución.

El archivo esta en formato FastA, por lo que tiene una línea con información, además las secuencias vienen en líneas de 80 base pairs (el salto de línea se cuenta como un caracter).

Vamos a contar el total de caracteres de todo el archivo que seria un aproximado del tamaño del genoma, o bien podemos sacar el valor exacto del tamaño del genoma al calcular la resta del total de carcateres del archivo menos el número de lineas y los caracteres del primer renglón.

> tamaño del genoma = total de caracteres - (num de lineas del archivo + numero de caracteres primera línea)

Para nuestro caso llegaremos hasta restar el número de líneas solamente.

```
wc -l -c data/coli_genomic.fna
```

Donde -l regresa el número de líneas, y -c el número de caracteres del archivo.

Resultado

La secuencia de nucleotidos del genoma completo de *E. coli* es de 4.6 MB, ya que el resultado del conteo es de 469,9745 caracteres del archivo, y restamos el numero de lineas 58022.

2. ¿Cuántos cromosomas tiene *Escherichia coli*?

La información de anotación de lo que se conoce en el genoma de *E. coli* se reporta en archivos de formato tabular GFF.

Archivo(s): data/coli.gff

> Como se describe en la sección `B. Datos de Entrada de éste documento, la columna 1 del formato GFF se anota el nombre del cromosoma donde esta reportada cada feature del genoma.
>
> El archivo tiene 7 lineas de comentarios (líneas que comienzan con `#`.
>
> El total de líneas del archivo son data/coli.gff que se obtiene con `wc -l data/coli.gff`

Algoritmo:

> - Usar la columna 1 del archivo GFF, que es el nombre del cromosoma.
>
>> - No nos interesa el resto de la información del `feature` solo en qué cromosoma esta anotado.
>> - Al quedarnos con el nombre del cromosoma, tendremos muchas repeticiones, tantas como `features` esten anotadas en ese cromosoma.
>> - Tomar en cuenta que las líneas de comentarios también vendrán.
>
> - Eliminar repeticiones (quedarnos con valores únicos de cromosomas)
> - Contar el número de cromosomas (Vendran las líneas de comentarios, hay que restarlas al total)

Solución

Siguiendo el algoritmo su implementación seria

```
cut -f1 data/coli.gff | uniq
```

donde `cut -f1` nos permite cortar la información de la columna 1, y `uniq` nos elimina repeticiones.

Resultados La respuesta que obtenemos es

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM584v2
#!genome-build-accession NCBI_Assembly:GCF_000005845.2
##sequence-region NC_000913.3 1 4641652
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=51114
NC_000913.3
###
```

Donde las líneas que comienzan con `#` son comentarios. Por lo que tenemos solo un cromosoma con identificador `NC_000913.3`.

Análisis y Conclusiones

El genoma de E.coli versión NC_000913.3, tiene ## cromosoma con un un tamaño aproximado de ### MB. Se han anotado ### número de genes, los cuales ### van en la cadena de 5'-3' y ### van en la cadena complementaria.

Referencias

[1] https://es.wikipedia.org/wiki/Escherichia_coli

[2] Frederick R. Blattner et al., The Complete Genome Sequence of *Escherichia coli* K-12.Science277,1453-1462(1997).DOI:10.1126/science.277.5331.1453