Practica01: El dogma central de la biología molecular

Equipo NewGen
Casas Espinosa, Axel;
Jimenez Reyes, Abraham; 318230577
Villarreal Maldonado, Jorge Manuel; 307312637

Genómica Computacional - 2024-1
Facultad de ciencias, Universidad Nacional Autónoma de México
22 de septiembre de 2023

En el estudio de muestras de ventilas hidrotermales, se analizaron fragmentos de DNA para identificar y entender la presencia de genes y posibles organismos que habitan estos ambientes extremos. Utilizando herramientas bioinformáticas, se leyeron secuencias en formato FASTA y se identificaron marcos de lectura abiertos (ORFs). Los ORFs se tradujeron en secuencias de aminoácidos para inferir funciones potenciales de las proteínas. Se determinó que uno de los fragmentos de DNA tenía una proteína relacionada con la formación del núcleo celular, sugiriendo la presencia de organismos eucariotas en la muestra. Esta revelación desafía el entendimiento tradicional de los tipos de vida que pueden soportar tales ambientes extremos. Adicionalmente, se discutió la complejidad de recrear organismos a partir de secuencias de DNA, destacando los desafíos técnicos y éticos, y el porqué aún no tenemos un "Jurassic Park.[en] la realidad.

Introducción

En este proyecto, estamos explorando la genómica de organismos que habitan en ambientes extremos, como las ventilas hidrotermales. Utilizando muestras recolectadas de estas áreas, hemos secuenciado fragmentos de ADN que podrían pertenecer a organismos desconocidos. El objetivo de este proyecto es analizar estas secuencias de ADN para identificar posibles genes y proteínas codificadas por estos fragmentos de ADN.

Marco teórico

Un *gen* es una unidad de herencia que ocupa una ubicación específica (locus) en un cromosoma. En el contexto de la biología molecular, un gen es una secuencia de nucleótidos en el ADN que codifica la síntesis de una cadena de polipéptidos o de una molécula de ácido ribonucleico (ARN) con una función conocida.

Un **Marco de Lectura Abierto** (ORF, por sus siglas en inglés, *Open Reading Frame*) es una secuencia continua de nucleótidos que tiene la potencialidad de codificar una proteína. Un ORF comienza con un codón de inicio (generalmente ATG, que codifica para el aminoácido metionina) y termina con uno de los tres codones de parada (TAA, TAG o TGA), sin ningún otro codón de parada en medio. Un ORF, por lo tanto, representa una parte de la secuencia de un gen que tiene el potencial de codificar una proteína. Para lograr este trabajo, se siguieron los siguientes pasos:

Parte 1: Descripción del Proyecto

Eres el computólogo a cargo en un equipo multidisciplinario donde tus colegas son biólogos, matemáticos y terrólogos. Resulta que tu grupo se dedica a investigar la vida que hay en ambientes extremos. Esta vez, lograron extraer muestras provenientes de ventilas hidrotermales. Los biólogos se hicieron cargo de su trabajo y te otorgaron esta carpeta con cuatro archivos FASTA.

Análisis de Búsqueda de Genes

Para realizar el análisis de búsqueda de genes en los archivos de secuencia, utilizamos una combinación de scripts de Bash y comandos de línea de comandos, incluyendo `grep` y `sed`. Para contar el número de genes (o encabezados de secuencia) presentes en cada archivo, utilizamos el comando `grep` con la opción `-c` para contar las líneas que comienzan con el carácter `>` (que denota el inicio de una nueva secuencia). Hay que tener en cuenta que esto se ejecuta en terminal, nosotros lo hicimos con Linux y tenemos la carpeta Venti-

lasHidrotermales con nuestros 4 archivos .fna.

```
grep -c "^>" *.fasta
  fragment_1.fna:1
  fragment_2.fna:1
  fragment_3.fna:1
  fragment_4.fna:1
```

Tambien lo podemos ejecutar de forma individual teniendo en cuenta que debemos ubicarnos a la altura de nuestros archivos .fna, esto quiere decir abrir nuestra terminal en la carpeta que tiene los archivos y escribir las siguientes lineas en la terminal.

```
grep -c ">" fragment_1.fna
grep -c ">" fragment_2.fna
grep -c ">" fragment_3.fna
grep -c ">" fragment_4.fna
```

Cálculo de la longitud de las secuencias

Para calcular la longitud de las secuencias en un archivo FASTA, podemos utilizar el comando `awk`.

```
awk '{if($0 !~ />/) print length($0)}'
file_name.fna
```

Hay que recordar que tenemos que tener los archivos .fna en la carpeta donde nos ubiquemos. Recordemos que colocamos una linea y presionamos enter, nos tiene que dar un numero diferente por cada archivo que es la longitud de la secuencia.

```
awk '{if($0 !~ />/) print length($0)}'
fragment_1.fna
awk '{if($0 !~ />/) print length($0)}'
fragment_2.fna
awk '{if($0 !~ />/) print length($0)}'
fragment_3.fna
awk '{if($0 !~ />/) print length($0)}'
fragment_4.fna
```

Análisis de Codones de Inicio y Parada

Para llevar a cabo un análisis inicial y localizar los codones de inicio y parada en las secuencias, podemos utilizar el comando `grep`. A continuación, presentamos una guía sobre cómo usar este comando para identificar líneas que contienen los codones de inicio (.ATG") y los codones de parada (”TAA", ”TAG", ”TGA") en un archivo FASTA:

Identificación de Codones de Inicio y Parada

- Codón de Inicio - **.ATG"**

  ```
  grep -n 'ATG' file_name.fna
  ```

- Codón de Parada - **”TAA"**

  ```
  grep -n 'TAA' file_name.fna
  ```

- Codón de Parada - ”TAG"

  ```
  grep -n 'TAG' file_name.fna
  ```

- Codón de Parada - ”TGA"

  ```
  grep -n 'TGA' file_name.fna
  ```

Identificación de Posibles Genes (ORFs)

```
grep -o -P 'ATG(?:[ATGC]{3}){30,}?T
(?:AA|AG|GA)' file_name.fna | wc -l
```

Explicación del Comando:
`grep`: Una herramienta para buscar cadenas específicas en archivos.

- `-o`: Opción que permite imprimir solo las partes de la línea que coinciden con el patrón.
- `-P`: Opción que permite interpretar la expresión como una expresión regular de Perl.
- `'ATG(?:[ATGC]3)30,?T(?:AA|AG|GA)'`: Es una expresión regular que busca una cadena.
- `file_name.fna`: Es el archivo donde se está realizando la búsqueda.
- `|`: Es un operador de tubería.
- `wc -l`: Cuenta el número de líneas en la entrada.

Escribimos cada linea en terminal para ver el numero de posibles genes en cada archivo .fna.

```
grep -o -P 'ATG(?:[ATGC]{3}){30,}?T(?:AA|AG|GA)'
fragmnet_1.fna | wc -l
grep -o -P 'ATG(?:[ATGC]{3}){30,}?T(?:AA|AG|GA)'
fragmnet_2.fna | wc -l
grep -o -P 'ATG(?:[ATGC]{3}){30,}?T(?:AA|AG|GA)'
fragmnet_3.fna | wc -l
grep -o -P 'ATG(?:[ATGC]{3}){30,}?T(?:AA|AG|GA)'
fragmnet_4.fna | wc -l
```

Los respectivos resultados de cada archivo .fna.

```
Para nuestro archivo fragment_!
encontramos 1309 genes.
Para nuestro archivo fragment_2
encontramos 208 genes.
Para nuestro archivo fragment_3
encontramos 1846 genes.
Para nuestro archivo fragment_4
encontramos 11 428 genes.
```

Utilizando BioPython para la Identificación de ORFs

En este proyecto, también nos apoyaremos en la biblioteca BioPython, una herramienta poderosa y flexible para el análisis computacional de secuencias biológicas. La biblioteca BioPython facilita la escritura de scripts de Python para trabajar con datos biológicos.

Instalación de BioPython

```
pip install biopython
```

Lectura de Archivos FASTA

Utilizaremos BioPython para leer los archivos FASTA y extraer las secuencias que contienen.

Identificación de ORFs

Posteriormente, escribiremos scripts para identificar ORFs en las secuencias extraídas. Esto implica buscar secuencias que comienzan con el codón de inicio .ªTGz terminan con un codón de parada (”TAA”, ”TAG.º ”TGA”), con una longitud mínima especificada para considerarse un ORF válido.

Código en Python

Este script de Python utiliza la biblioteca BioPython para analizar archivos en formato FASTA. Identifica y traduce los marcos de lectura abiertos (ORFs) en las secuencias nucleotídicas presentes en el archivo. Los ORFs se definen como secuencias que comienzan con el codón .ªTGz terminan con uno de los codones de parada (”TAA”, ”TAG.º ”TGA”), con una longitud mínima de 90 nucleótidos (30 tripletes). El script también ofrece detalles sobre la longitud de cada secuencia y el número de ORFs identificados. El codigo lo ejecutaremos de la siguiente manera, abriremos nuestra terminal y nos posicionaremos en la carpeta que tenga nuestro codigo.Escribiremos en nuestra terminal la siguiente linea sin comillas

```
p1_NewGen.py
```

En nuesto codigo tenemos algo llamado file_path que es la direccion de nuestros archivos .fna. Tenedremos que colocar la direccion de nuestros archivos que estan en la carpeta llamada VentilasHidrotermales. Si surge algun error basta con mover cada archivo .fna a la altura de nuestro codigo (omitir la carpeta VentilasHidrotermales) y en el file path colocaremos solo el nombre del archivo por ejemplo

```
file_path = fragment_1.fna
```

es en la linea 30 de nuestro codigo y cambiamos el numero para cada archivo. Enseguida se muestran los resultados.

```
1  from Bio import SeqIO
2  from Bio.Seq import Seq
3
4  def read_fasta(file_path):
5      sequences = {}
6      for record in SeqIO.parse(
           file_path, "fasta"):
7          sequences[record.id] = str(
               record.seq)
8      return sequences
9
10 def find_orfs(sequence, min_length
       =100):
11     orfs = []
12     sequence_length = len(sequence)
13     for i in range(sequence_length):
14         if sequence[i:i + 3] == 'ATG'
               :
15             for j in range(i,
                   sequence_length, 3):
16                 if sequence[j:j + 3]
                       in {'TAA', 'TAG',
                       'TGA'}:
17                     if j + 3 - i >=
                           min_length:
18                         orfs.append(
                               sequence[i
                               :j + 3])
19                     break
20     return orfs
21
22 def translate_orf(orf):
23     return Seq(orf).translate(to_stop
           =True)
24
25 file_path = "./fasta/fragment_1.fna"
26 sequences = read_fasta(file_path)
27
28 for seq_id, seq in sequences.items():
29     print(f"ID␣de␣la␣Secuencia:␣{
           seq_id},␣Longitud:␣{len(seq)}"
           )
30
31 orfs_in_sequences = {}
32 for seq_id, seq in sequences.items():
33     orfs_in_sequences[seq_id] =
           find_orfs(seq)
```

```
34
35 for seq_id, orfs in orfs_in_sequences
       .items():
36     print(f"ID␣de␣la␣Secuencia:␣{
           seq_id},␣N mero␣de␣ORFs:␣{len
           (orfs)}")
37     translated_orfs = [translate_orf(
           orf) for orf in orfs]
```

Resultados (Parte 1)

Análisis de Fragmentos de Secuencia

Archivo Fragmento 1:

- Número de registros (posibles genes): 1
- Detalles del registro:
  - ID: Fragmento1
  - Longitud de la secuencia: 632,428 bases
  - Número de ORFs identificados: 3,611

Archivo Fragmento 2:

- Número de registros (posibles genes): 1
- Detalles del registro:
  - ID: Fragmento2
  - Longitud de la secuencia: 100,531 bases
  - Número de ORFs identificados: 527

Archivo Fragmento 3:

- Número de registros (posibles genes): 1
- Detalles del registro:
  - ID: Fragmento3
  - Longitud de la secuencia: 1,154,456 bases
  - Número de ORFs identificados: 5,741

Archivo Fragmento 3:

- Número de registros (posibles genes): 1
- Detalles del registro:
  - ID: Fragmento4
  - Longitud de la secuencia:7,115,445 bases
  - Número de ORFs identificados: 35,429

Observaciones:

- **Fragmento 1:** Aunque su tamaño sugiere que podría ser un genoma bacteriano, el número de ORFs identificados (3,611) es típico para muchas bacterias. Esto refuerza la idea de que este fragmento puede ser un genoma bacteriano completo.
- **Fragmento 2:** A pesar de su menor tamaño, el número de ORFs (527) es notable. Aunque no es suficiente para ser un genoma bacteriano completo, podría representar un plásmido o un genoma viral.
- **Fragmento 3:** Con 5,741 ORFs, este fragmento tiene una diversidad genética considerable. Esto, junto con su longitud, sugiere que es muy probable que sea un genoma bacteriano completo.
- **Fragmento 4:** La enorme cantidad de ORFs (35,429) refuerza la idea de que este fragmento pertenece a un organismo eucariota, como un hongo o incluso un protista.

Parte 2: Identificación y Extracción de un Gen Clave en Ambientes Extremos

Tus colegas descubren que el primer gen del fragmento de DNA más corto, podría ser importante para los organismos que viven en ventilas hidrotermales.

Código Python (GenTranslator)

Para compilar este codigo necesitaremos que nuestro archivo GenTranslator.py este en la misma carpeta que nuestros archivos .fna ya que para ejecutarlo con exito necesitamos escribir en la linea 73 lo siguiente

```
seq = read_first_sequence_from_fasta
("fragment_4.fna")
```

Esto por que nos generara 3 archivos .fasta, el nombre es el mismo para cada archivo .fna entonces si queremos los 3 archivos para el fragment_1 los tenemos que copiar o guardar ya que al cambiar al fragment_2 genera los mismos 3 archivos pero es la informacion de este archivo .fna. Teniendo en cuenta lo anterior, abriremos nuestra terminal y escribiremos

```
p1_2NewGen.py
```

damos enter y verificamos en nuestra carpeta que se crearon los 3 archivos.

El programa `Gen Translator` se diseñó para analizar y procesar genes de interés en secuencias de DNA. A partir de un archivo `.fasta` con una secuencia genética, el programa realiza las siguientes tareas:

- **Generación del cDNA**: Calcula la secuencia complementaria de DNA a partir de la secuencia de entrada y la guarda en un archivo `cDNA.fasta`.
- **Transcripción a ARNm**: Convierte la secuencia de DNA en su correspondiente ARN mensajero, eliminando timinas y reemplazándolas por uracilos. El resultado se almacena en `mRNA.fasta`.
- **Traducción a Aminoácidos**: Traduce el ARN mensajero en una cadena de aminoácidos utilizando el código genético, generando una secuencia proteica. Esta secuencia se guarda en `aminoacidos.fasta`.

Resultado (Parte 2)

A partir de la información genómica y proteómica obtenida y con base en los archivos generados (`cDNA.fasta`, `mRNA.fasta`, `aminoacidos.fasta`), hemos deducido que el organismo en estudio es eucarionte. A continuación, se detallan las razones principales:

Gen Relacionado con la Formación del Núcleo Celular

Se identificó un gen que está asociado con la formación o función del núcleo celular. Es esencial destacar que solo las células eucariontes poseen un núcleo celular bien definido. La presencia de este gen es una fuerte evidencia de la naturaleza eucariota del organismo.

cDNA y su Relevancia en Eucariontes

El archivo `cDNA.fasta` representa el ADN complementario formado a partir del ARNm. En eucariontes, el cDNA es crucial para estudiar la expresión génica, ya que refleja exclusivamente los genes expresados, excluyendo las regiones intrónicas. Los procariotas carecen de intrones, por lo que la importancia del cDNA en este contexto sugiere un origen eucarionte.

Genes en el ADN: Exones e Intrones Los genes en el ADN están compuestos por dos tipos principales de secuencias: exones e intrones.

Exones Son las secuencias de ADN que se transcriben y traducen en proteínas. Es decir, tienen la información codificada que se utilizará para producir una proteína específica.

Intrones Son las secuencias de ADN que se encuentran entre los exones, pero no se traducen en proteínas. Durante el proceso de formación del ARN mensajero (ARNm) en eucariotas, los intrones se transcriben inicialmente, pero luego son eliminados en un proceso llamado .empalme.o ”splicing”, dejando solo los exones en el ARNm maduro.

Características del ARN Mensajero

El archivo `mRNA.fasta` contiene secuencias de ARNm. En eucariontes, este ARNm pasa por un proceso de maduración que incluye adiciones específicas y el empalme para eliminar intrones. Estas características, si se detectan en el archivo, indican un proceso típico de maduración del ARNm eucarionte.

Análisis Proteómico

El archivo `aminoacidos.fasta` presenta las proteínas traducidas a partir del ARNm. Un análisis posterior de estas secuencias reveló funciones específicas asociadas a eucariontes.

Ensayo (Parte3)

Pregunta genuina: Si el DNA contiene toda la información que nos conforma como organismos vivos, como especie y como organismos únicos, ¿Por qué no existe *Jurassic Park/World*? ¿Es fácil crear un organismo en el laboratorio solo con tener una cadena DNA?

¿Por qué aún no tenemos un Jurassic Park?

¡Hola! Si eres de los que crecieron viendo ”Jurassic Parkz soñando con ver un Tiranosaurios rex en vivo y en directo, te comprendo totalmente. Pero, aunque la idea de recrear dinosaurios y pasear entre ellos suena alucinante, la ciencia real detrás de esto es un poco más complicada que insertar un poco de ADN antiguo en un huevo y esperar a que salga un velociraptor. Así que, antes de que empieces a ahorrar para tu entrada al parque de los dinosaurios, aclaremos un par de cosas.

Primero, el ADN. Sí, es cierto que el ADN es como el manual de instrucciones que nos hace ser lo que somos. Piénsalo como ese complicado manual con el que armas un mueble. Pero, en lugar de tornillos y tablones, el ADN tiene las instrucciones para construir un ser vivo, ya sea una bacteria, un girasol, tú o un brontosaurio.

Ahora, imagina que encuentras un viejo manual, pero la mitad de las páginas están dañadas o faltan. Eso es lo que pasa con el ADN de dinosaurios. Aunque hemos encontrado ámbar con mosquitos que chuparon la sangre de dinosaurios (¡sí, como en la película!), el ADN en él está superdañado. Y, para empeorar las cosas, el ADN no dura eternamente. Después de un tiempo, se rompe en pedacitos, y después de millones de años, bueno... es más difícil de encontrar que un calcetín.

Entonces, aunque tengamos fragmentos de ADN de dinosaurio, reconstruir todo el manual es como intentar armar el mueble con solo un tercio de las instrucciones y sin saber qué mueble es.

Pero, supongamos que, mágicamente, logramos tener todo el ADN. Ahí no terminan los problemas. Para que un dinosaurio nazca, necesitas mucho más que solo su ADN. Es como tener el manual, pero sin las herramientas, tornillos o incluso el espacio adecuado para armarlo. Necesitaríamos un óvulo de dinosaurio (que, obviamente, no tenemos) y una máquina del tiempo para traer a una madre dinosaurio dispuesta a incubar el huevo.

Ahora, hablemos de la parte de çrear organismos en el laboratorio". En teoría, con el ADN correcto y las herramientas adecuadas, podríamos hacerlo. De hecho, ya estamos experimentando con cosas como ovejas clonadas y edición genética. Pero de ahí a recrear un organismo extinto hay un gran trecho. Y, honestamente, aunque pudiéramos... ¿Deberíamos? Un mundo con dinosaurios suena genial en el cine, pero en la vida real, habría un montón de problemas éticos y prácticos. ¿Dónde vivirían? ¿Cómo los alimentaríamos? Y, lo más importante, ¿cómo nos aseguramos de que no se coman a los turistas?

En resumen, aunque la idea de un Jurassic Park es emocionante, la ciencia detrás de ello es complicada y, por ahora, está más en el terreno de la ciencia ficción que en la realidad. Así que, por el momento, lo más cercano que tendremos a un dinosaurio será ese pollo rostizado del domingo. Después de todo, ¡los pájaros son descendientes directos de los dinosaurios! Y, si piensas en ello, eso es igual de impresionante.