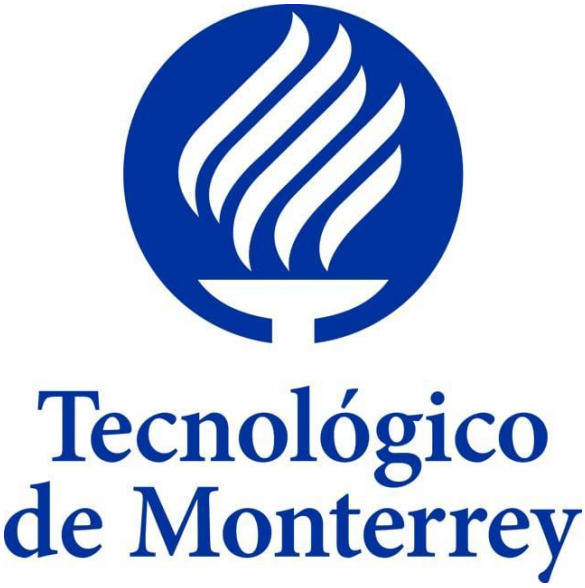
Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey



Análisis y diseño de algoritmos avanzados
Dra. María Valentina Narváez Terán

Actividad Integradora 1

Alonso Abimael Morales Reyna - A01284747
Diego Alejandro Hernandez Romero A01198079
Jose Eduardo Gomez Saldaña A00833968
Melilssa Elvia Salazar Carrillo A01383422
Santiago Andrés Gámez Campos A01197653
Sergio Ortíz Malpica A01284951

Grupo 608

6/10/23

Actividad Integradora

Link al Replit: https://replit.com/join/exeyltujpo-a01284747

link video:https://drive.google.com/file/d/1y7IdViq06mz0mocpWFjcnn8Y_8B0n2nj/view?usp=sharing

Problema y Soluciónes

El problema que queremos resolver es llegar a conocer más sobre la secuencia del SARS-COV-2 de Wuhan, el virus que causó la pandemia de 2020. Para conocer más esta secuencia necesitamos encontrar el índice donde se encuentran tres genes del virus, al igual que los palíndromos más largos de los tres. Queremos encontrar los palíndromos debido a que son áreas propensas a mutaciones. Después tenemos que encontrar el índice de ocurrencia de las proteínas que tiene el virus. Finalmente, requerimos comparar el genoma del virus de Wuhan de 2019 con uno encontrado en Texas en 2020 con el propósito de encontrar dónde difieren y si es que estos cambios resultan en aminoácidos diferentes.

Para encontrar los índices de aparición de cada uno de los tres genes decidimos utilizar el algoritmo **Knuth-Morris-Pratt (KMP)** debido a que es un algoritmo sencillo de implementar y tiene una complejidad temporal de **O(n+m)**. Para encontrar el índice de ocurrencia también utiliza una función auxiliar que genera una tabla **LPS (Longest Prefix Suffix)** para encontrar el prefijo más largo que también es el sufijo más largo del patrón a partir de cada índice. Esta tabla se utiliza para determinar dónde resumir una búsqueda cuando se produce una discrepancia entre el texto y el patrón que se quiere encontrar.

Se utilizó el algoritmo de **Manacher** para encontrar el palíndromo más largo de cada uno de los tres genes. Las estructuras de datos que se utilizaron fueron los siguientes:

- Cadena Transformada (**T**): Es una versión de la cadena original, pero con caracteres especiales añadidos. Facilita el manejo de palíndromos de ambas longitudes, par e impar.
- Array de Palíndromos (**P**): Almacena la longitud del palíndromo más largo centrado en cada posición **T**.

Decidimos elegir este algoritmo porque ofrece una solución más rápida y eficiente. A diferencia de los enfoques convencionales operan en O(n^2), Manacher lo logra en **O(n)**. La eficiencia proviene de la manera en cual se reutiliza la información previamente calculada sobre palíndromos para evitar recalculaciones innecesarias. Esta rapidez es crucial para secuencias genéticas largas.

Para encontrar cuáles secciones del virus se produce cada proteína, regresando su nombre, sus índices, sus cuatro primeros aminoácidos y los codones asociados, utilizamos el algoritmo **Algoritmo Z**, dado que nos puede ayudar para encontrar esto, y que puede ser bastante eficiente.

Para el último punto debíamos encontrar diferencias entre los genomas del virus Wuhan 2019 vs Texas 2020, mostrar los codones afectados por esas diferencias y el aminoácido producido por cada codón, para este problema fue esencial la estructura de datos tipo diccionario de python ya que esta nos facilitó la interpretación de los codones a aminoácidos de una manera eficaz y rápida, para buscar las diferencias reutilizamos principios del algoritmo **KMP** y fuimos comparando los virus en subcadenas de longitud 3 para de esta manera saber cómo se modifica el codón dependiendo el cambio, ya que la diferencia era un solo carácter en una secuencia inmensa por lo cual se debía dividir, almacenamos los índices donde se presentaban estas diferencias en una lista, esto junto con el codón correspondiente y su definición de aminoácido.

En cuanto a las estructuras de datos usadas, se utilizan las siguientes:

- Cadenas de texto: Se utiliza una cadena de texto para almacenar el string con el carácter para buscar.
- Array: se utiliza el array o lista para guardar los valores de Z en este.
- Diccionario: Se implementó un diccionario para poder obtener los codones asociados a cada aminoácido.
- Listas: Se implementaron listas para almacenar de manera sencilla datos obtenidos

En este caso, la complejidad del programa es de **O(n * m),** dado que en este caso, se cuenta con un recorrido adicional para encontrar los aminoácidos y sus codones.

Creemos que esta complejidad es de buena eficiencia comprando con otros métodos, aunque existe potencial de mejorarla, tanto en la implementación del algoritmo como la lectura del archivo.

Reflexiones

Alonso

Con esta actividad pude ver cómo lo que he aprendido en esta clase se puede aplicar en la vida real en áreas de la ciencia que no están relacionadas a mi carrera. Me di cuenta de la importancia que tienen los algoritmos para ayudar a encontrar soluciones a problemas complejos. En este caso, analizar una secuencia de un virus para encontrar dónde es más propenso a mutar y analizar las diferencias entre dos variaciones de un virus. Con esta actividad tuve algunos problemas para hacer que un algoritmo funcione, por lo que decidí probar otro para ver que resultado obtenía, lo cuál me hizo ver que en ocasiones es bueno probar varios algoritmos para ver cuál es el más adecuado para resolver un problema de manera más eficiente.

Sergio

En esta clase de evidencias se puede apreciar la importancia de la implementación de esta clase de algoritmos en ramas muy diferentes a lo que es una ingeniería. Uno pensaría que algoritmos de este tipo comúnmente se utilizan en desarrollo de aplicaciones con motores de búsqueda por ejemplo, que es un tema más orientado hacia la carrera de ingeniero computacional. Sin embargo para esta evidencia la utilizamos en un aspecto muy diferente. Los algoritmos que analizan strings juegan un papel fundamental en la sociedad gracias a su capacidad de encontrar patrones en un texto haciendo posible que analizen incluso palabras. Algo similar sucede con las tecnologías de inteligencia artificial que analizan patrones en los textos para responder al usuario. Para este caso utilizamos esta clase de algoritmos para analizar secuencias de adn del virus sars cov, de modo que pudiéramos romper la secuencia del virus en secciones obteniendo los datos que queremos de la secuencia de adn del mismo virus.

En lo personal me gustó más trabajar con el algoritmo KMP, el cual fue utilizado para el primer punto para poder encontrar en qué índices de la secuencia de adn del virus se encontraban ciertos genes. Intentamos este punto con un algoritmo de z array sin embargo se nos facilitó y resultó más eficiente utilizar el KMP. No llegue a dimensionar que la programación y algoritmos son muy importantes también para el mundo de la medicina incluso pero ahora puedo dimensionar y darme cuenta de la importancia que tienen esta clase de algoritmos. Me agrado esta actividad, puse en práctica lo aprendido en clase sobre algoritmos de búsqueda de patrones y logre comprender la importancia que pueden llegar a tener en la sociedad.

Diego

Está situación problema dejó muchos aprendizajes, él mas importante a mi parecer es él como un problema en está caso de la vida real puede ser resuelto por un conjunto de algoritmos computacionales de manera eficiente y además no usando demasiados recursos, él conocer estos algoritmos nos da la ventaja de poder implementarlos en un futuro en otros problemas en vez de tratar de resolverlos por medio de un método ineficiente como lo pudiera ser la fuerza bruta. Es increíble ver como usando algoritmos en clase se pueden llegar a hacer cosas como estas las cuales si pudieran llegar a tener un gran impacto en él mundo real.

Jose

Esta primera situación aunque generó un reto grande para el equipo me dejó muchos aprendizajes para entender mejor cómo manejar cadenas de caracteres lo cual es muy útil en distintos problemas que se pueden presentar, además conocimos más a fondo algoritmos como KMP y Z array lo cual facilitaron la búsqueda de patrones dentro d e una cadena de texto y gracias a conocer el verdadero funcionamiento de estos algoritmos pudimos implementar soluciones para buscar coincidencias entre dos cadenas de texto y no solo buscar un patrón en específico. Aprendí la importancia de escoger de manera correcta la estructura de datos para el almacenamiento y acceso de

los datos, ya que si no hubiéramos usado diccionarios en algunos casos se pudiera haber complicado más tal vez con una matriz, etc.

Melissa

Al concluir con esta situación problema, no sólo aprendí la implementación técnica, sino también sobre la importancia del pensamiento algorítmico en la solución de problemas en el mundo actual. A través del análisis y comparación de diferentes algoritmos, puede apreciar cómo cada uno tiene sus propias ventajas y limitaciones. Esta actividad me ayudó a adentrarme más en la intersección de la informática con la biología. Fue interesante ver como conceptos teóricos cobraban vida y relevancia en un contexto real. Aprendí a manejar mejor mis tiempos, a analizar qué actividades son prioridad en comparación con otras. Es gratificante ver el resultado de nuestro trabajo.

Santiago

Esta actividad me ha ayudado a entender mejor el uso de los algoritmos que se utilizan para manejar y analizar strings, y de igual manera su importancia. Ahora que utilizamos los diferentes algoritmos para cada punto, fue evidente ver que el trabajo de analizar es mucho más eficiente y menos tardío que si se hiciera a mano y sin uso de estas tecnologías. Esto es de alto valor al ahorrar tiempo y recursos necesarios, y en el caso de trabajo con viruses, ayuda a que el tiempo que se requiere para entender cómo funcionan los viruses y como es que están compuestos sea menor, y asi, que la mayoria del enfoque sea entorno a encontrar curas para estos. Esto igual ayuda muchas diferentes áreas, tanto profesional (podría servir para hacer revisiones de texto extremadamente largos y encontrar puntos de interés, como errores, elementos relevantes, etc) como personal (ayuda a eficientar búsquedas en línea y en páginas web, al igual que ayudar en dar herramientas que mejorar la experiencia de usuario). Aunque fue un poco retador al tener que familiarizarnos con algunos conocimientos biológicos para cumplir los objetivos, y de igual manera adaptar lo que hemos estado trabajando para esta situación problema, esto se pudo atender y nos ayudó a seguir reafirmando lo que sabemos.

Capturas de pantalla

```
1. Encontrar el índice de aparición de cada uno de los genes en la secuencia
gen-M
El gen se encuentra en el índice 26523 del virus SARS-COV-2 de Wuhan
Los primeros 12 caracteres del gen son: ATGGCAGATTCC

gen-S
El gen se encuentra en el índice 21563 del virus SARS-COV-2 de Wuhan
Los primeros 12 caracteres del gen son: ATGTTTGTTTTT

gen-ORF1AB
El gen se encuentra en el índice 266 del virus SARS-COV-2 de Wuhan
Los primeros 12 caracteres del gen son: ATGGAGAGCCTT
```

```
2. Encontrar el palíndromo mas largo en cada uno de los tres genes
El palíndromo más largo en el gen M es:  CTAAAGAAATC
El palíndromo más largo en el gen S es:  ATTTTGGTGGTTTTA
El palíndromo más largo en el gen ORF1AB es:  CTCAATGACTTCAGTAACTC
```

main.py × *palindromos_mas_largos.txt* × +

palindromos_mas_largos.txt

```
1  Gen M:
2  Palíndromo: CTAAAGAAATC
3  Longitud: 11
4
5  Gen S:
6  Palíndromo: ATTTTGGTGGTTTTA
7  Longitud: 15
8
9  Gen ORF1AB:
10 Palíndromo: CTCAATGACTTCAGTAACTC
11 Longitud: 20
12
13
```

```
4. Diferencias Wuhan 2019 vs Texas 2020:
Índice: 8781, Codon en Wuhan 2019: TCA Aminoacido: S, Codon en Texas 2020: CCA Aminoacido: P
Índice: 19173, Codon en Wuhan 2019: GCT Aminoacido: A, Codon en Texas 2020: GAT Aminoacido: D
Índice: 27924, Codon en Wuhan 2019: TAA Aminoacido: *, Codon en Texas 2020: CAA Aminoacido: Q
Índice: 28143, Codon en Wuhan 2019: CAC Aminoacido: H, Codon en Texas 2020: TAC Aminoacido: Y
Índice: 29094, Codon en Wuhan 2019: TGG Aminoacido: W, Codon en Texas 2020: CGG Aminoacido: R
Índice: 29880, Codon en Wuhan 2019: AA Aminoacido: ?, Codon en Texas 2020: AAA Aminoacido: K
```