Alineamiento de secuencias genética con álgebra lineal

Nicolás Gómez Aragón*

Pontificia Universidad Javeriana

November 9, 2023

Abstract

El propósito de este proyecto es presentar un enfoque simplificado del algoritmo de alineamiento de secuencias de ADN utilizando conceptos de álgebra lineal. Se explicará cómo las matrices pueden utilizarse para encontrar similitudes entre secuencias de ADN. Además, se proporcionará un ejemplo de alineamiento de secuencias en Python y se demostrará en clase. Esto permitirá comprender cómo se aplican las herramientas matemáticas vistas en clase en el campo de la genética y la bioinformática.
Palabras Clave: Genética, Álgebra Lineal, Alineamiento, secuencias, bioinformática.

Contents

1 Objetivos del Proyecto

- Introducir la bioinformática con el tema de secuencias genéticas.
- Explicar el uso de matrices en el alineamiento de secuencias genéticas.
- Presentar el algoritmo Needleman-Wunsch de manera simplificada.
- Proporcionar un ejemplo práctico de alineamiento de secuencias en Python.

*nicolas.gomeza@javeriana.edu.co

2 Introducción

A pesar de que la rama científica que estudia los genes, sus variaciones y la herencia en organismos puede considerarse como uno de los campos más recientes de la biología, la genética se ha estudiado de manera extensa y con proyectos de alto impacto científico. Desde las memorias de Gregor Mendel en 1865 hasta el Proyecto del Genoma Humano en los años 90´s y finalizado en 2003, con los avances tecnológicos y la implementación de métodos matemáticos para solucionar problemas, el campo de la genética se ha convertido de gran interés para la comunidad científica.[1]

Algo muy importante para el estudio de la genética han sido las matemáticas, que junto con las computadoras, han sido herramientas poderosas para estudiar amplios volúmenes de datos de manera eficiente y precisa. Desde la herencia[2], el álgebra genética [3] o algoritmos genéticos [1], el uso particular del álgebra ha tenido una importancia fundamental en el estudio de la genética.

3 Preliminares

El ADN, también llamado ácido desoxirribonucleico, es una molécula fundamental en biología que lleva la información genética necesaria para el desarrollo, funcionamiento y reproducción de los seres vivos. Es reconocida por su estructura de doble hélice, y es una de las piezas fundamentales de la vida.

Una secuencia de ADN es una cadena de moléculas compuesta por cuatro bases nitrogenadas diferentes, y estas son: Adenina (A), Timina (T), Citosina (C) y Guanina (G). Estas contiene la información genética que define las características y funciones de un organismo, son únicas para cada individuo y son fundamental para la síntesis de proteínas y la regulación de funciones celulares. Dentro de una secuencia genética, cada letra representa una base nitrogenada que forma parte del ADN. [4]

El alineamiento de secuencias implica la comparación de secuencias de proteínas o ácidos nucléicos (cadena molecular) para revelar similitudes y diferencias, esto permite identificar regiones con patrones coincidentes y es fundamental para tareas como la búsqueda de genes, la comprensión de la evolución genética y la identificación de variantes genéticas. Además, el alineamiento de estructuras tridimensionales es crucial para analizar la similitud en la forma de moléculas biológicas.[5]

4 Algoritmos para el Alineamiento Genético

Un algoritmo es una serie de pasos que se siguen para realizar un proceso y con ello obtener algún resultado, en este caso, el algoritmo de alineamiento de secuencias de ADN nos permite encontrar similitudes entre secuencias genéticas y revela regiones de coincidencia y diferencias. Estos algoritmos son esenciales para tareas como la búsqueda de genes, la comprensión de la evolución genética y la identificación de variantes genéticas en volúmenes grandes datos eficientemente.

Los algoritmos de alineamiento genético son una clase de herramientas computacionales que permiten comparar secuencias biológicas para identificar regiones de similitud y entender su estructura y función. Estos algoritmos desempeñan un papel crucial en la interpretación de datos genéticos y han revolucionado la genómica y la biología molecular. A medida que la cantidad de datos genéticos disponibles ha aumentado exponencialmente, la necesidad de algoritmos eficientes y precisos de alineamiento genético se ha vuelto cada vez más evidente.En el vasto campo de la bioinformática, donde se entrelazan la genética, la informática y las matemáticas, se encuentran diversas técnicas destinadas a la comparación de secuencias genéticas.

En el contexto de esta exposición, abordaremos un algoritmo de notable relevancia: el **Algoritmo Needleman-Wunsch**. Este destaca por su influencia duradera y su enfoque basado en la programación dinámica. Este algoritmo, concebido por Saul Needleman y Christian Wunsch en 1969, sienta las bases para la comparación precisa de secuencias genéticas.[6]

El legado del Algoritmo Needleman-Wunsch se extiende más allá de la investigación académica y ha permeado la práctica clínica, la genómica y la biología molecular. Ha permitido identificar genes conservados, descubrir relaciones evolutivas y analizar genomas completos. Su aplicación ha impulsado avances significativos en campos como la medicina personalizada, el descubrimiento de fármacos y la ingeniería genética.[5]

En las próxima sección, exploraremos en detalle el funcionamiento del algoritmo y cómo se aplica.

5 Algoritmo Needleman-Wunsch: Fundamentos

El algoritmo Needleman-Wunsch permite encontrar la alineación óptima de dos secuencias genéticas al asignar puntuaciones a las coincidencias entre las secuencias y las diferencias entre los elementos genéticos.

5.1 Programación Dinámica en Alineamiento de Secuencias

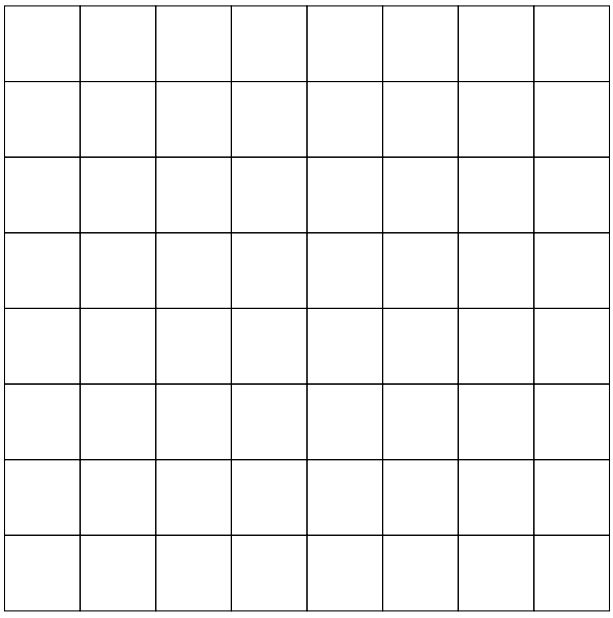
El Algoritmo Needleman-Wunsch utiliza programación dinámica, una técnica de optimización que descompone un problema en partes más pequeñas y las resuelve de manera eficiente. En el contexto del alineamiento de secuencias genéticas, aplica este enfoque para encontrar la solución global, dividiendo el proceso en segmentos más manejables. De esta forma, puede comparar secuencias de aminoácidos o ácidos nucleicos de diferentes longitudes y detectar similitudes, incluso cuando hay inserciones o eliminaciones en las secuencias. Las inserciones y eliminaciones se refieren a la introducción de espacios (gaps) o elementos en una secuencia para alinearla con la otra o la remoción de elementos. Estas operaciones son esenciales para ajustar las secuencias y lograr el mejor alineamiento posible. Asimismo, el algoritmo utiliza matrices de puntuación, donde cada elemento representa la similitud entre pares de bases en las secuencias, siendo fundamentales para calcular la mejor alineación que maximiza la puntuación total.

5.2 Matriz de puntuación

De acuerdo a esto, podemos tomar 2 cadenas de nucleótidos, dado que es más simple que con una proteína por su simplicidad química, así:

$$x = CAGCTA,\ y = CACATA$$

Este es un problema bidimensional, donde podemos crear una matriz de las dimensiones de las cadena genéticas que vamos a comparar y adicionarle dos, en este caso, una matriz dos cadenas genéticas de 6 dimensiones nos llevarían a crear una matriz de 8x8:

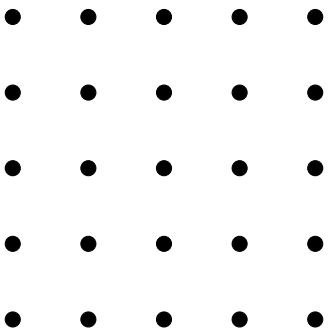


Con esta malla que previamente construimos, podemos ver que el algoritmo propone tomar dos secuencia de n dimensiones y colocar una de manera horizontal, representando el eje x y otra vertical representando el eje y, así cada combinación de dos letras sería una celda. Por consiguiente, la puntuación para las celdas adyacentes se calcula con base en los puntajes que serán descritos más adelante.

A esta matriz, la llamamos matriz de puntuación, y cada celda representa la mejor alineación entre dos subsecuencias. Para llenar la matriz puede iniciarse desde la esquina superior izquierda, hasta la esquina inferior derecha o al revés, y se va avanzando aplicando el sistema de puntuación de manera recursiva.

5.3 Problema de Optimización y Grafos

Ya con esto tenemos las bases del problema, con lo cuál nuestro objetivo es buscar un camino a través de esta matriz, el cual maximice la puntuación, es decir, maximizar la puntuación de aciertos entre nuestras dos cadenas genéticas, de esta forma podríamos ver el problema de la siguiente manera:



Así, el problema de optimización también podría interpretarse como un problema de grafos, específicamente un grafo dirigido, donde tenemos una estructura conformada por vértices, unida por aristas y se busca el camino que acierte la mayor puntuación. Podríamos afirmar que es un grafo dirigido dado que cada arista tiene una orientación y es acíclico, ya que el nodo de inicio no es el que termina el ciclo trazado.

Esta interpretación de la situación se suele llamar "Problema del Viajante de Comercio" (TSP por sus siglas en inglés, Traveling Salesman Problem), este es un problema clásico de optimización de grafos en el que se busca determinar la ruta más corta que visita un conjunto de ciudades exactamente una vez y regresa al punto de partida. El TSP es un problema NP-duro y ha sido ampliamente estudiado en la teoría de la computación y la optimización combinatoria, mostrando como el algoritmo también podría aplicarse en otros ámbitos. [5]

5.4 Cálculo de Puntuación

De esta forma, podemos definir una matriz de sustitución que llevará el registro de puntuaciones para cada iteración. En esta matriz, cada elemento representa la comparación de una base nitrogenada de una cadena genética (x) con la correspondiente en otra cadena (y). De esta manera, podemos calcular y comparar las puntuaciones de alineación entre las dos secuencias genéticas dependiendo de nuestra ubicación:

$$S_{x,y} = \begin{cases} 1 & \text{si } x = y \\ -1 & \text{si } x \neq y \end{cases} \tag{1}$$

La puntuación de cada celda está determinada por la puntuación de las celdas adyacentes, y se utiliza una función, como la expuesta previamente (1), para determinar la puntuación de la alineación. Este sería un ejemplo de nuestra matriz de sustitución:

$$S_{a,b} = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

5.5 Penalizaciones

Cuando los elementos en una secuencia genética no coinciden, se considera un espacio, lo que indica que algo se ha insertado o eliminado en las secuencias para lograr la concordancia. Esta representación se conoce como penalización o penalidad, ya que implica un costo algorítmico por los ajustes necesarios en las secuencias genéticas para que los elementos concuerden. Esta penalización, comúnmente denominada "penalización de espacio lineal", es un valor predeterminado que se resta del puntaje de alineación. Cuanto mayor sea la penalización, menor será el puntaje total al final de la alineación de las secuencias generadas por el algoritmo. Podemos definir esto como:

$$c_L(d) = d \cdot G$$

lo cual es directamente proporcional a la longitud d del espacio por un parámetro $G < 0$.

Existen otros tipos de penalizaciones como la "penalización de espacio afinado", las cuáles penalizan abrir un espacio y expandirlo de manera diferente, lo cual va más allá del alcance de este artículo.[7]

5.6 Recursión en el Algoritmo

Con esta restricción, se adiciona otra capa de complejidad al algoritmo Needleman-Wunsch, donde no solo se cuentan elementos que concuerden o no, sino también el costo de introducir nuevos espacios. Así con la configuración de la matriz y la función definida, se puede proceder a determinar la puntuación máxima de la alineación a través de la matriz.[6] De este modo, podemos definir cada paso como un proceso recursivo para definir la ruta más óptima:

$F_{i-1,j-i}$	$F_{i-1,j}$
$F_{i,j-1}$	$F_{i,j}$

Con esto, el proceso de recursión se vería de la siguiente manera en forma de función:

$$F_{i,j} = \begin{cases} F_{i-j,j-1} + S_{x[i],y[j]} & \text{si acierto/desacierto} \\ F_{i-1,j} + G & \text{si espacio en y} \\ F_{i,j-1} + G & \text{si espacio en x} \end{cases} \tag{2}$$

Existe un sitio web, el cuál demuestra de manera simple lo que se ha buscado explicar en este artículo. Para visitarlo, puede acceder en este enlace.

6 Algoritmo Needleman-Wunsch en Python

A pesar a haber explicado el algoritmo con aparente simplicidad, su implementación en código es más compleja de lo que podría parecer y fue tomado de un repositorio de un tercero para poder mostrarlo con efectividad.

Así, el algoritmo debidamente comentado se encuentra en el siguiente repositorio.

7 Conclusiones

En este proyecto, nos hemos adentrado en el fascinante mundo del alineamiento de secuencias de ADN, una rama en la genética y la bioinformática que desempeña un papel crucial en estas. Nos centramos en el Algoritmo Needleman-Wunsch, una herramienta poderosa que utiliza conceptos de álgebra lineal para calcular similitudes entre secuencias genéticas y encontrar alineamientos óptimos. A través de este enfoque, pudimos destacar la importancia de las matrices en la resolución de desafíos genómica.

El proyecto también subraya la influencia significativa del álgebra lineal en el campo de la genética, demostrando cómo las herramientas matemáticas desempeñan un papel fundamental en la comprensión y el análisis de datos genéticos. El Algoritmo Needleman-Wunsch, basado en la programación dinámica y matrices de puntuación, emerge como un pilar de la medicina personalizada, la genómica y la ingeniería genética. Su aplicación va más allá de este campo, y puede tener diversas aplicaciones en otros campos académicos.

A manera personal, me entusiasmó el proceso de investigación y experimentación para realizar el proyecto, y me sorprende como la implementación en código python era prácticamente la traducción del desarrollo matemático que realizamos en este documento, lo cual nuevamente dme muestra el poder y versatilidad de las matemáticas.

References

Nicholas J. Radcliffe. The algebra of genetic algorithms. *Annals of Mathematics and Artificial Intelligence*, 10:339–384, 1994.

Jared Kirkham. Linear algebra applications in genetics. page 6, 2001.

Bernard Russo. Evolution algebra. Presentation, 2015. Freshman Seminar, University of California, Irvine.

Encyclopaedia Britannica. *dna*, 2023. Last Updated: Oct 22, 2023.

Phillip Compeau. *Bioinformatics Algorithms: An Active Learning Approach*. Active Learning Publishers, La Jolla, CA, 2015.

Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

Henry Hendrix. Understanding sequence alignment algorithms: with needleman-wunsch, 2022.