Nicole M. Ramirez Mulero (801-19-3940)

Prof. Humberto Ortiz

CCOM 6685

5 de mayo de 2025

Alineación de distancia para secuencias de AND y ARN

I. Abstracto

Comparar secuencias es importante para muchas aplicaciones en la vida real, como identificar anormalidades o mutaciones en secuencias, e identificar zonas biológicamente significativas [1]. Para esta comparación Podemos utilizar alineamiento por distancia. El alineamiento de edición implica encontrar el mínimo número de ediciones con tal de que dos secuencias estén lo más alineadas posible. Para lograr este alineamiento por distancia utilizamos el algoritmo de Needleman-Wunsch, algoritmo usado en la bioinformática para alinear secuencias de proteínas o nucleótidos. Este algoritmo es una manera inteligente para procesar todas las posibilidades de secuencias que se deben considerar. Reduce significativamente el tiempo requerido y garantiza una solución óptima.

II. Introducción

La comparación de dos secuencias en el ámbito de la biología tiene muchas implementaciones importantes para los avances científicos [1]. Cuando se compran 2 secuencias esto permite descifrar códigos biológicos complejos. En el caso del ADN o del ARN la comparativa de secuencias permite entender como estas codifican secuencias de aminoácidos [2]. La comparativa también se usa para casos de identificación de anormalidades en las secuencias

de ADN, ayudando a la detección de ciertos tipos de cáncer [2]. Estas comparaciones también ayudan a identificar regiones conservadas biológicamente significativas. Las cadenas de ADN pueden ser extremadamente largas, pero se han identificados áreas cortas altamente similares y significativas. La comparación de secuencias permite encontrar estas áreas locales [2].

Diversas herramientas se han creado a base de la comparación de secuencias para realizar tareas fundamentales en la bioinformática y la genómica. La distancia por edición calcula el número de operaciones mínimo-necesarias para transformar una secuencia en otra, lo que indica el porcentaje de diferencia que existe entre ambas secuencias. Alineamiento solapado se refiere a encontrar esas áreas dentro de una secuencia que se superpone, y este alineamiento ayuda al a samblaje de ganonas a partir de fragmentos. El Alineamiento de ajuste compara una secuencia corta con una más larga con tal de encontrar el mejor ajuste posible [1].

En el caso de este reporte, el tipo de alineamiento que estaremos utilizando es el alineamiento por edición. El alineamiento de edición implica encontrar el mínimo número de ediciones con tal de que 2 secuencias estén lo más alineadas posible. Esto implica el ordenamiento de un par de secuencias genéticas de AND o ARN para determinar las regiones de similitudes o diferencia. Su objetivo es determinar el alineamiento optimo con la puntuación más alta [3]. Es decir, el máximo nuevo de coincidencias base a base, sin alterar el orden de las bases en ninguna de las secuencias.

III. El algoritmo: Needleman-Wunsch

Durante la clase de Bioinformática tuvimos la oportunidad de implementar el LCSBacktrack(), una matriz de alineamiento de secuencias de ADN que mostraba la matriz de alineamiento para 2 secuencias. Esta matriz muestra una flecha vertical si ningún cambio era

necesario, una línea vertical para letras incorrectas y una línea horizontal para espacios entre la secuencia. Esta implementación poseía fallas ya que solo contaba las coincidencias entre secuencias, llevando a insertar espacios erróneos. Para reparar esos, implementamos el alineamiento local, que también añade una penalización para espacios y desajustes. El problema que se resolvió en este reporte es, a base de esta matriz de alineamiento global, crear las secuencias alineadas, incluyendo los espacios necesarios para acercar lo más posible estas 2 secuencias. Para lograr esto se estuvo implementando el algoritmo Needleman-Wunsch, un algoritmo usando en la bioinformática para alinear secuencias de proteína o nucleótidos [2].

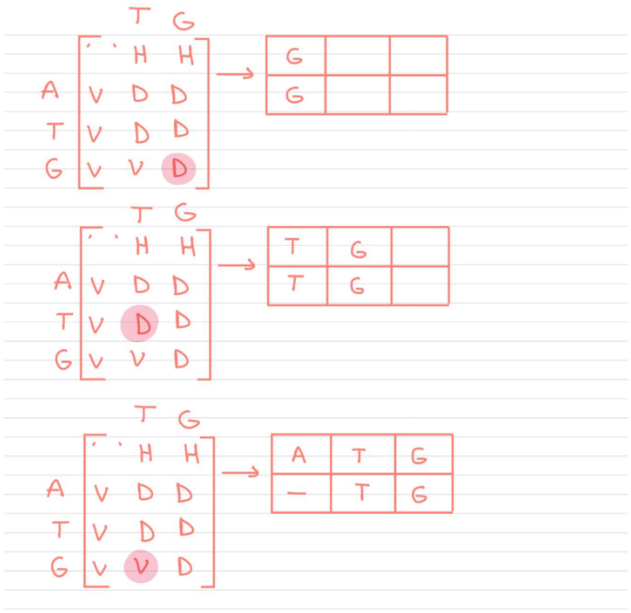
Este algoritmo implementado de manera “bruta” (brute-force) es un algoritmo que tomaría demasiado tiempo y recursos. Su complejidad de tiempo seria exponencial, lo que no lo hace efectivo [1]. Para solucionar esto se le puede aplicar un método llamado programación dinámica. La programación dinámica es una técnica en la que un problema se divide en más problemas, se guardan los resultados y luego se optimizan los subproblemas para encontrar la solución general [3]. Esta solución normalmente tiene que ver con encontrar el rango máximo o mínimo.

El algoritmo de Needelman-Wunsch se basa en una matriz de alineamiento y en una serie de puntajes. Si los caracteres son iguales, la puntuación será +1. Si los caracteres no son iguales, la puntuación es de -1. Y si es necesario un espacio, ya sea porque la adición de un espacio hace que las secuencias este más alineadas o que las secuencias no sean del mismo tamaño, la puntuación es +2. Estas puntuaciones varían por el tipo de secuencia. Cada tipo de secuencia requiere una puntuación distinta ya que hay secuencias que son más susceptibles que otras. En el caso de las cadenas de ADN y ARN se requiere una puntuación de -2 ya que las secuencias de ADN y ARN son secuencias que se basan en 4 (A, C, T, G) caracteres por ende no es tan

susceptible como una secuencia que use más caracteres [4]. En el área de pruebas se demuestra esto.

Estas puntuaciones se utilizan en la matriz. A cada celda de la matriz se le hace 3 cálculos. Primero el cálculo diagonal, donde se le suma +1 al valor en diagonal superior si los caracteres son iguales, o -1 si no lo son. Luego se calcula de valor horizontal (valor a la izquierda de la celda que estamos calculado) y su valor vertical (valor arriba de la celda que estamos calculando) y a cada uno se le suma -2. De estos 3 valores se eligen el valor máximo y es el valor que se pone en la celda. Al mismo tiempo se crea otra matriz que va a funcionar como el seguimiento. Dependiendo de le valor que se escoja (ya sea diagonal, vertical y horizontal) se va a colocar en esa celta el tipo de valor que se utilizó. “D” si se utilizó el valor diagonal, “H” si se usó el valor horizontal y “V” si se utilizan el valor vertical. A continuación, se muestra un ejemplo.

Para construir las secuencias ya alineadas usamos principalmente la matriz de seguimiento. Esta Matriz se va a leer de atrás hacia delante de derecha a izquierda en diagonal y los caracteres se añaden comenzando con el primero hasta el último. Si el valor en la cerda es D, significa que los caracteres son iguales y solo se pasan a la secuencia final, si es “V” significa que un espacio debe de ser incluido en la segunda secuencia y si es “H” significa que un espacio debe de ser incluida en la primera secuencia. Un ejemplo se muestra a continuación.

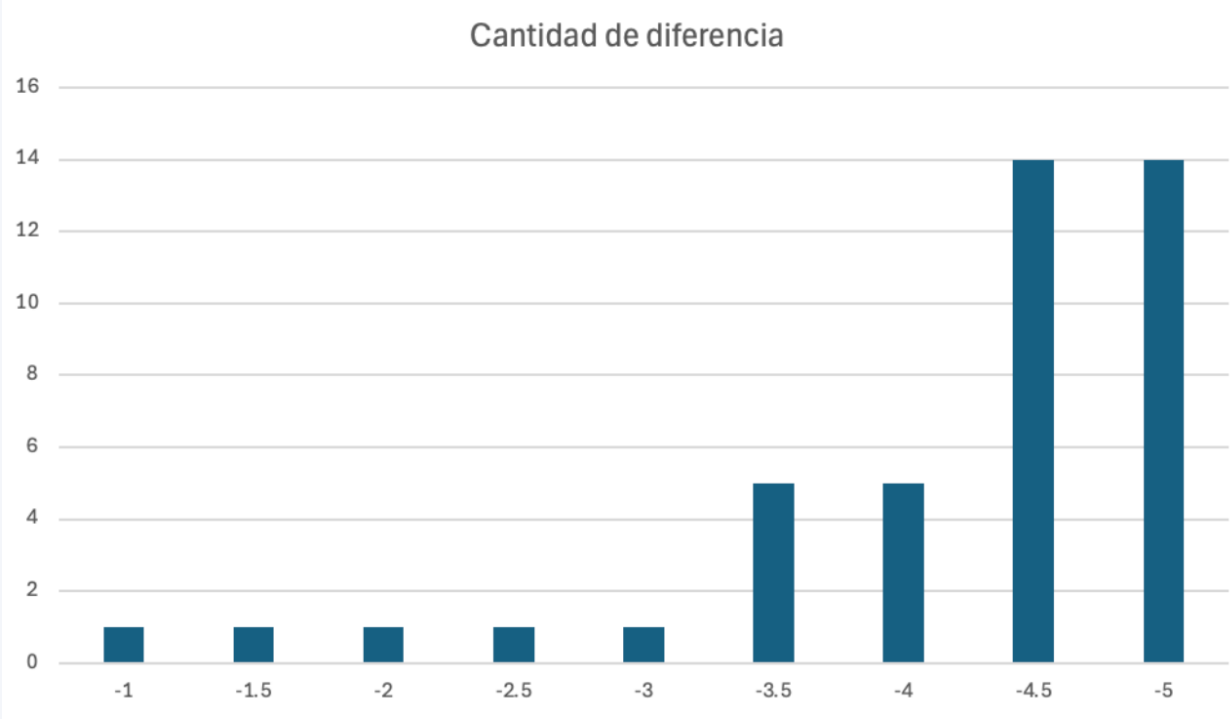


El código para este algoritmo puede ser encontrado en la página de github: https://github.com/Nicole-M-Ramirez/Needleman-Wunsch

III. Resultados

Para las pruebas primero se realizaron pruebas normales para comprimir que el código alineara bien secuencias de distintos tamaños. Estas pruebas fueron bien realizadas por el programa. Otro tipo de pruebas que se le realizaron al programa fue para verificar que puntaje de los espacios era óptimo para alinear las secuencias de la mejor manera posible. Esto se calculó

teniendo puntajes de -1 a -5 en intervalos de -0.5. Los resultados revelaron que desde -1 a -3 no hay errores substanciales en las secuencias alineadas, pero ya desde -3.5 empiezan a haber errores no esperados en las secuencias y ya en-4.5 los errores son substanciales. Cabe recalcar que no todas las secuencias son igual de susceptibles, incluiso siendo de AND o ARN. Hay secuencias que incluso con un puntaje de -10 se alinean correctamente. Pero para darle el mejor oportunidad a todas las secuencias se utiliza un valor medio entre -1 y -3, en nuestro caso -2.



IV. Conclusión

En la Bioinformática, comparar secuencias es crucial. Utilizando el algoritmo de Needleman-Wunsch implementado con programación dinámica, fue posible crear alineaciones óptimas para cadenas de AND y ARN, independiente de la longitud de la secuencia y cuan susceptible esta sea. Esta metodología nos ayuda a identificar similitudes significativas y diferencias notables, un aspecto crucial para identificar mutaciones, analizar áreas preservadas y respaldar diagnósticos

moleculares. Los resultados de este estudio subrayan la importancia de ajustar con exactitud las sanciones por espacios vacíos y errores, dado que una calibración equivocada podría poner en riesgo la calidad de la alineación. Para las secuencias de ADN y ARN, se ha demostrado que una penalización de -2 es un valor compensado que funciona adecuadamente.

V. Bibliografía

[1] Compeau, P., & Pevzner, P. (2018). Bioinformatics algorithms: an active learning approach. La Jolla, California: Active Learning Publishers. ISBN-13978-0990374633 https://www.bioinformaticsalgorithms.org/

[2] Likic, V. (2008). The Needleman-Wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne, 1-46.

[3] Rashed, A. E. E. D., Amer, H. M., El-Seddek, M., & Moustafa, H. E. D. (2021). Sequence alignment using machine learning-based needleman–wunsch algorithm. IEEE Access, 9, 109522-109535.

[4] Kenneth H. Rosen. 2010. Handbook of Discrete and Combinatorial Mathematics, Second Edition (2nd. ed.). Chapman & Hall/CRC.