**Estructura de datos y algoritmos**

**Taller 3**

**Ejercicio 2.** *Distancia de Hamming.*

**2.** Dé la complejidad en tiempo y la complejidad espacial del algoritmo.

*La complejidad en tiempo del algoritmo es donde es la longitud de las secuencias. Esto se debe a que nuestra función "distancia\_hamming" itera sobre las secuencias para comparar los caracteres en la misma posición. Esto significa que el tiempo de ejecución del algoritmo aumenta linealmente según el tamaño de las secuencias.*

**3.** ¿Cuál es el principal inconveniente del uso de esta distancia en el contexto de secuencias ADN?

*El principal inconveniente del uso de la distancia de Hamming en el contexto de secuencias de ADN radica en su limitación para capturar las variaciones genéticas que incluyen inserciones, eliminaciones y sustituciones de nucleótidos. Aunque la distancia de Hamming puede ser efectiva para medir la diferencia entre secuencias de igual longitud, no considera las operaciones biológicas anteriormente mencionadas, que son fundamentales en el análisis de la evolución y similitud de las secuencias de ADN. Por lo tanto, su aplicación en este contexto puede resultar insuficiente. En lugar de utilizar la distancia de Hamming, se prefieren métodos más avanzados como el análisis de k-mers mencionado en el contexto proporcionado, ya que pueden abordar de manera más completa la complejidad de las secuencias de ADN y proporcionar una comprensión más precisa de su similitud.*

**Ejercicio 3.** *Descomposición en k-mers y función hash.*

**2.** ¿Cuántos *k*-mers se generan por una secuencia de longitud ?

*Para una secuencia de longitud y un valor de k, se generarán k-mers. Esto se debe a que cada k-mer es una subsecuencia de longitud k que se puede obtener desplazando una ventana de tamaño k a lo largo de la secuencia. Por lo tanto, hay posibles desplazamientos.*

**3.** ¿ Cuál es la complejidad en tiempo del algoritmo de descomposición en *k*-mers de una secuencia *S* dada ?

*La complejidad en tiempo del algoritmo de descomposición en k-mers de una secuencia dada es , donde es la longitud de la secuencia y es el tamaño de los k-mers. Esto se debe a que se debe iterar sobre la secuencia una vez, generando un k-mer en cada iteración. Como se mencionó anteriormente, hay posibles desplazamientos para generar k-mers en una secuencia de longitud . Por lo tanto, la complejidad es lineal respecto a la longitud de la secuencia y al tamaño de los k-mers.*

**Ejercicio 4.** *Similitud genética y índice de Jaccard.*

Dé los resultados sobre el conjunto de datos dado en forma de matriz de similitud para K = 7 y K = 11. Indique el tiempo de ejecución para cada K.

|  |  |  |
| --- | --- | --- |
| dataset\k | 7 | 11 |
| dataset xylella | [[1. 0.99536896 0.99554987 ... 0.01062466 0.01441045 0.01471576]  [0.99536896 1. 0.99530545 ... 0.01062985 0.0144175 0.01472295]  [0.99554987 0.99530545 1. ... 0.010637 0.01442719 0.01473285]  ...  [0.01062466 0.01062985 0.010637 ... 1. 0.01234568 0.00484262]  [0.01441045 0.0144175 0.01442719 ... 0.01234568 1. 0.00632911]  [0.01471576 0.01472295 0.01473285 ... 0.00484262 0.00632911 1. ]]  Tiempo de ejecución para k = 7: 73.64888954162598 segundos | [[1.00000000e+00 5.99844695e-02 4.04476830e-02 ... 9.22319420e-05  7.44768005e-05 1.34776147e-04]  [5.99844695e-02 1.00000000e+00 3.61456876e-02 ... 5.48215558e-05  1.56004267e-04 1.68656106e-04]  [4.04476830e-02 3.61456876e-02 1.00000000e+00 ... 4.00701673e-05  1.06830473e-04 5.34126214e-05]  ...  [9.22319420e-05 5.48215558e-05 4.00701673e-05 ... 1.00000000e+00  0.00000000e+00 0.00000000e+00]  [7.44768005e-05 1.56004267e-04 1.06830473e-04 ... 0.00000000e+00  1.00000000e+00 0.00000000e+00]  [1.34776147e-04 1.68656106e-04 5.34126214e-05 ... 0.00000000e+00  0.00000000e+00 1.00000000e+00]]  Tiempo de ejecución para k = 11: 132.85821723937988 segundos |