Práctica 2

Indexación de secuencias

Objetivo

Conocer estructuras de datos y algoritmos para la compresión e indexación de secuencias.

Tareas

A partir de un fragmento de 20 bases de una secuencia de ADN real (obtenida por ejemplo de https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi):

1. (0,15 puntos) Construir el array de sufijos correspondiente a dicho fragmento.
2. (0,15 puntos) Seleccionar un patrón cualquiera de 3 caracteres e indicar cómo se obtendrían las posiciones de sus ocurrencias en el fragmento.
3. (0,15 puntos) Construir las estructuras del array de sufijos comprimido de Sadakane, indicando los valores de las estructuras de Ψ, D y S.
4. (0,15 puntos) Usando sólo Ψ, D y S, indica cómo se obtendrían los primeros 4 caracteres del fragmento de ADN utilizado.
5. (0,15 puntos) Usando sólo Ψ, D y S, indicar cómo se contarían cuántas ocurrencias existen del patrón elegido en la tarea 1.
6. (0,1 puntos) Obtener la transformada de Burrows-Wheeler.
7. (0,15 puntos) Para el patrón elegido en la tarea 1, indicar cómo se obtendría el número de ocurrencias en el texto utilizando *backward search*.

Entregables

Un archivo comprimido con el siguiente contenido:

1) Un documento pdf con la solución a las tareas propuestas. Se debe indicar claramente qué fragmento de secuencia de ADN y qué patrón se ha utilizado.

Fecha de entrega

Para obtener la puntuación completa se debe entregar a través de la plataforma campus vritual antes de las 23:55 del 16 de abril de 2021.

Se podrá entregar con penalización (80% de la nota) hasta el 7 de mayo de 2021.