

Proyecto 2

Sistemas Operativos

Segundo Semestre 2023, Prof. Cecilia Hernández

Fecha Inicio: Lunes 30 de Octubre, 2023.

Fecha Entrega: Viernes 24 de Noviembre, 2023 (23:59 hrs).

Trabajo en grupo: Integrado con 3 estudiantes. Una entrega por grupo.

Entrega: Archivo comprimido con readme.txt e informe.

1. Objetivos

- Desarrollar habilidades de programación multihebra y mecanismos de sincronización en los estudiantes.

2. Metodología: Trabajo en grupo de 3 estudiantes.

3. Descripción

Suponga existe un sistema donde tiene una colección muy grande de archivos que desea procesar para extraer ciertas características de los datos y solo desea almacenar la información para solo el conjunto de archivos que cumple con ciertas restricciones.

En particular considere una colección de genomas, los cuales son archivos de texto que contienen las cuatro bases nitrogenadas A, C, T, y G. Para el procesamiento se desea obtener el contenido genético CG de un genoma, el cual consiste en la proporción de C y G respecto del total de bases, es decir está dado por $\frac{C+G}{C+G+A+T}$. Se desea seleccionar los genomas con contenido GC superior a un umbral dado.

Considere la colección de genomas que está disponible en el link genomas y los subdirectorios que se encuentran allí con nombres bacterias, samples y invertebrates. Para procesar los genomas debe ignorar las líneas que comienzan con caracter.

Para resolver el problema debe usar múltiples hebras para procesar los genomas y una cola compartida para guardar la información del genoma y contenido GC que supere un umbral dado.

a) Requerimientos

- 1) Su implementación debe realizarla en C/C++ y linux.
- 2) Su aplicación debe recibir como parámetro el directorio entrada donde se encuentra la colección de genomas y el umbral de contenido genético CG.
- 3) Investigue como leer archivos de un directorio dado.
- 4) Utilice múltiples hebras leer la colección de archivos, donde cada una debe procesar cada genoma y escribir en cola compartida.
- 5) Implemente la cola compartida usando semáforos por las hebras que permita almacenar solo los nombres de genomas con su contenido genético que supere un umbral.
- 6) Implemente la cola compartida usando mutex y variables de condición por las hebras que permita almacenar solo los nombres de genomas con su contenido genético que supere un umbral.

- 7) Su implementación de cola compartida debe contener un método que permita imprimir los genomas con contenido genético GC mayor al umbral definido.
- 8) Realice un análisis comparativo en tiempo y espacio de memoria requerido entre las dos implementaciones. Para medir el tiempo y espacio de memoria mediante el Maximum Resident Set. Para simplificar esta evaluación baste que use el comando `time` de linux (no el de la shell), para lo cual deberá averiguar como usarlo correctamente.
- 9) Incluya un `readme.txt` y un informe en pdf describiendo su implementación y su evaluación experimental.