

Proyecto Final: Análisis de Rendimiento de Dotplot Secuencial vs. Paralelización

Descripción General

El objetivo de este proyecto es implementar y analizar el rendimiento de tres formas de realizar un dotplot, una técnica comúnmente utilizada en bioinformática para comparar secuencias de ADN o proteínas.

Los estudiantes implementarán una versión secuencial, una versión con hilos, una versión paralela utilizando la biblioteca *multiprocessing* de Python, una versión paralela utilizando *mpi4py*. Compararán el rendimiento de estas tres implementaciones utilizando varias métricas y visualizarán sus resultados.

Requisitos del Proyecto

Implementación

1. Aplicación de línea de comandos para dotplot: Los estudiantes deben crear una aplicación de línea de comandos que pueda tomar dos secuencias (en formato de archivos FASTA) como entrada y producir un dotplot de las secuencias. La aplicación debe ser capaz de ejecutar el dotplot de tres maneras: secuencialmente, utilizando *multiprocessing*, utilizando *mpi4py*.

Análisis de rendimiento

Los estudiantes deben calcular y analizar las siguientes métricas para cada implementación:

1. Tiempos de ejecución totales y parciales (porción paralelizable).
2. Tiempo de carga de los datos y de generación de la imagen.
3. Tiempo muerto (tiempo no empleado en la ejecución del problema).
4. Aceleración y eficiencia.
5. Escalabilidad.

Se deben realizar gráficas de desempeño, aceleración, eficiencia y escalabilidad para facilitar la comparación entre las tres implementaciones.

Datos de prueba

Los estudiantes deben probar su aplicación con un archivo FASTA que contiene dos secuencias de bacterias. Estos archivos proporcionarán un buen conjunto de datos de prueba debido a su tamaño y las expectativas de similitud entre las secuencias.

Entrega

El proyecto final debe ser entregado en un repositorio de GitHub que incluya el código fuente de la aplicación, los archivos de datos de prueba, y el informe con los resultados. El repositorio también debe incluir un archivo README que describa cómo ejecutar la aplicación y cualquier otra información que sea útil para entender el proyecto.

Evaluación

El proyecto será evaluado según los siguientes criterios:

- Implementación del dotplot.
- Cobertura y precisión del análisis de rendimiento.
- Estrategia de paralelización realizada.

Recursos:

- Carpeta con los scripts y recursos vistos en clase : [Carpeta de la Sesión en clase](#)
- Archivos con los que deben probar la aplicación: [Insumos \(archivos para ejecutar el programa\)](#)
- Artículo científico de G-SAIP, un programa para realizar un dotplot: [Paper de G-SAIP](#)
- Repositorio de G-SAIP (para que se guíen en la elaboración de la aplicación por línea de comandos): [Repo: G-SAIP: Graphical sequence alignment with HPC](#)

Recuerden que la aplicación debe ser ejecutable por línea de comandos. Ej:

```
mpirun -n 4 python miAplicacionMPI.py - -file1=salmonella.fna - -file2=E.coli.fna --thres=0.8  
-- output=miDotplot.py
```

¡¡Muchos éxitos!!