CAB401   
  
PARALLELIZATION PROJECT REPORT

Dac Duy Anh Nguyen n10603280

Promoter

Table of Contents

[I. Original sequential application 3](#_Toc84884666)

[1) What it does 3](#_Toc84884667)

[2) How it works 3](#_Toc84884668)

[II. Potential parallelism analysis 3](#_Toc84884669)

[III. Tools and techniques utilized 4](#_Toc84884670)

[IV. Outcomes 4](#_Toc84884671)

[V. Difficulties 4](#_Toc84884672)

[VI. Reflection 4](#_Toc84884673)

# I. Original sequential application

## 1) What it does

The application is utilized to identify different promoters in reference list from Ecoli bacteria

## 2) How it works

Firstly, the program loads Ecoli DNA and reference genes using the available helper functions.

In the **first** **For** **loop** (number 1), the program iterates through each Ecoli DNA file to retrieve its gene record.

Another **inner For loop** (number 2) iterates through the reference genes which will then be compared with the each Ecoli gene in the following **inner For loop** (number 3).

The comparation process is handled by Homologous function to determine if 2 genes serve the same purpose. In this case, they are a gene from Ecoli records (from For loop 3) and another from reference genes (from For loop 2).

If the Homologous function return true, the program proceeds to extract upstream region of the Ecoli gene and use it to predict if that Ecoli gene is a promoter. If true, the results are stored in ‘consensus’ variable by increase the counter of that promoter type and total matches by 1 each.

\*\*\*Watch intro video again

# II. Potential parallelism analysis

|  |
| --- |
| Your analysis of potential parallelism within the application. This might include  identification of existing loops or control flow constructure where parallelism might be found. Explanation of the data and control dependences that you analyzed to determine which sections of code were safe to parallelize. Which of these is likely to be of sufficient granularity to be worth exploiting? Is it scalable parallelism? A discussion of changes required to expose parallelism, such as replacing algorithms or code restructuring transformations.  c. map computation to processor  It uses 3-nested for loop which runs and analyzes large text files (DNA) and is being executed in a sequential way. As a consequence, the computing time takes a substantial amount of time and could be reduced if applying parallelization. |

Initial analysis:

Transforming the For loops structure: move For loop 3 to inner For loop 1 and execute For loop 3 before For loop 2. However, extraction and prediction parts will remain in For loop 2. This makes the memory and processes relating to Ecoli genes data are close to each other, locality?

After the transformation above, the program can be paralleled at each of the For loops.

Identified dependencies:

* upStreamRegion
* prediction
* consensus

# III. Tools and techniques utilized

Hardware specs:

Software:

Techniques:

Parallel stream

ThreadPool

ExecutorService

Locality

# IV. Outcomes

d. timing and profiling results, speedup graphs

e. same results (number of promoters identified)?

# V. Difficulties

Identify data dependencies and make some variables ThreadLocal, avoid data race

Speedup after parallelizing 3 for loops is still not good. Identify causes…

# VI. Reflection