CAB401   
  
PARALLELIZATION PROJECT REPORT

Dac Duy Anh Nguyen n10603280

Promoter

Table of Contents

[1. Original sequential application 3](#_Toc85756017)

[1.1 What it does 3](#_Toc85756018)

[1.2 How it works 3](#_Toc85756019)

[2. Potential parallelism analysis 4](#_Toc85756020)

[3. Tools and techniques utilized 5](#_Toc85756021)

[3.1 Hardware specs: 5](#_Toc85756022)

[3.2 Software: 5](#_Toc85756023)

[3.3 Techniques: 5](#_Toc85756024)

[3.3.1 Explicit Threading 6](#_Toc85756025)

[3.3.2 Parallel stream 6](#_Toc85756026)

[3.3.3 Executor Service 6](#_Toc85756027)

[4. Outcomes 7](#_Toc85756028)

[5. Difficulties 7](#_Toc85756029)

[6. Reflection 7](#_Toc85756030)

[7. References 7](#_Toc85756031)

[8. Appendixes 7](#_Toc85756032)

# 1. Original sequential application

## 1.1 What it does

The application is utilized to identify different promoters – whose genes are listed in reference list – from the Ecoli bacteria genes.

For more information regarding the biology in this app, referenceGenes.list contains 8 pairs of gene name and its gene - the protein sequence. Each letter in the sequence corresponds to an amino acid – a basic building block of a protein sequence.

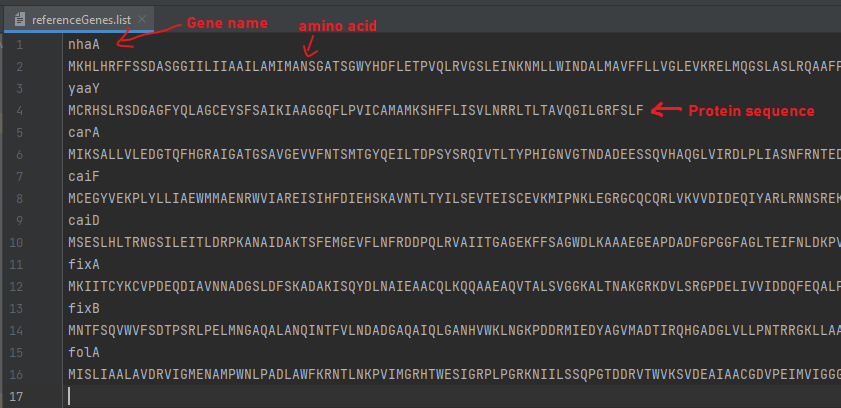


Figure Capture of the referenceGens.list file

In each gene bank file (with .gbk tail), ORIGIN section is the DNA of a bacteria organism. The DNA can contain millions of DNA molecules represented by 4 letter a-t-g-c and triplet of DNA molecules encodes 1 protein. There are 20 different amino acids out of 64 possible combinations of a triplet, and they are the basic building block of proteins. Some sections of DNA don’t correspond directly to proteins so in between proteins, there are upstream region called promoters which are the objects we need to identify.

## 1.2 How it works

### 1.2.1 Important packages

Description and analysis of some important packages:

* SmithWatermanGotoh: contains utility classes that uses Smith-Waterman algorithm with Gotoh’s improvement for biological local pairwise sequence alignment. Note that classes in this package requires an intensive computation power to determine the similar regions between two strings of sequences.
* Sigma70Consensus: initializes the map and prediction of a sigma factor 70 consensus and contains utility functions to record results.
* BioPatterns: contains getBestMatch function which is responsible for searching the best match of a pattern in a nucleotide sequence.
* GenBankRecord: its instance contains a list of gene names, their strand, location and gene sequences and a nucleotide sequence. This package also has Parse method used to extract the mentioned data from a file.

### 1.2.2 Sequential

Through a brief analysis in run() function, the program first loads Ecoli DNA and reference genes using ParseReferenceGenes function.

After that, in the **first** **For** **loop** (number 1), the program iterates through each Ecoli DNA file to retrieve its gene record using Parse function.

Another **inner For loop** (number 2) iterates through the reference genes to compare it with each Ecoli gene in its **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 an Ecoli gene from Ecoli records (from For loop 3) and another gene from reference genes (from For loop 2).

Inside the Homologous function, gene sequences are passed to align function of the SmithWatermanGotoh package.

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

# 2. Potential parallelism analysis

## 2.1 Profiler analysis

From the initial analysis of the given sequential program, Homologous function dominantly consumes 99.7% of the CPU computation. More specifically, it’s the align function in SmithWatermanGotoh class that get called by the Homologous function.

Text

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Since SmithWatermanGotoh function is a complicated algorithm and required a deep understanding of problem to rewrite for better performance, the work of this report mainly focuses to parallelize the app and increase performance at other parts where possible.

More importantly, this app uses 3-nested for loop which is being executed sequentially and includes many computations, read and write operations inside the for loops. As a consequence, the computing time takes a substantial amount of time and could be reduced if applying parallelization.

## 2.2 Dependencies:

sigma70\_pattern is assigned value from Sigma70Definition.getSeriesAll\_Unanchored(). But the method itself getSeriesAll\_Unanchored() is also static which will cause data race when there are multiple threads trying to access sigma70\_pattern. So to parallelize, I made it a ThreadLocal variable to store data individually for each thread (except for Explicit Threading technique).

Control dependency: the condition statement in ‘if’ affects whether the code in it executes or not. However, the properties in Gene class are not static meaning values of its instance are separated from each thread so it can be safe to implement parallelization for this part.

A lot of data dependencies are also identified in the most inner section of the For loops. Specifically, they are of True Dependence type in which one statement reads a value written by an earlier statement. In this case, they are upStreamRegion and prediction.

Text

Description automatically generated

Figure 2 Control (red) and data dependencies (orange)

# 3. Tools and techniques utilized

## 3.1 Hardware specs:

* CPU: Intel Core i7-8565U @1.8GHz
* 4 cores
* 8 virtual cores
* 8GB RAM
* Cache
  + L1 Data 4x32KB, 8-way
  + L1 Instruction 4x32KB, 8-way
  + L2 4x256KB, 4-way
  + L3 8MB, 16-way

## 3.2 Software:

- Integrated development environment: IntelliJ

- Java 15

- Run on Window Operating System

- Profiler tool: Jprofiler

## 3.3 Techniques:

In this project, I restructured the For loops code block to parallelize (except for Explicit threading technique). I made For loop 2 wrap For loop 1 and 3. Now, the order of the 3 For loops looks like this:

2: referenceGene -> 1:filename -> 3:gene

Text

Description automatically generated

Figure 3 For loops initial order

Text

Description automatically generated

Figure 4 For loops order changed

The aim of this transformation is to increase spatial locality by reducing gap time between accessing ‘record’ in For loop 1 and 3 and therefore, increase the number of cache hits for ‘record’ when transiting from For loop 1 to 3

While this puts referenceGene further from the inner part (*if (Homologous(…){ …}*), the high frequency in calling the variable will keep its memory location stored in cache.

There are 3 parallelization approaches I have implemented in this project:

* Explicit Threading
* Parallel Stream
* Executor Service

### 3.3.1 Explicit Threading

This is the first method I attempted for this application. From the sequential code, I thought the application can run faster by parallelizing and letting each thread handles one single input file instead of having them queue and running one by one. Therefore, I estimated this method can reduce the execution time by approximately 4 times (with 4 threads running parallelly to process 4 given Ecoli text files at once).

To apply explicit threading method, I created a new class called ExplicitThreading so that it can spawn a new thread of Sequential class by a separated For loop. To distribute the workload, that For loop also pre-processes ListGenBankFiles and assign each file for each thread. With this change, the Sequential class now only handle one file in each thread and 4 of them can run simultaneously.

With this approach, there are several changes in the original code structure. In order to prepare and extract the necessary data and add them as parameters to each Sequential call, 2 functions, ProcessDir and ListGenBankFiles, from Sequential class were moved down to my newly created class.

Text

Description automatically generated

Figure 5 Newly created ExplicitThreading Class

New private variables and constructors were also added in Sequential to keep exclusive data from ExplicitThreading for each particular thread.

Text

Description automatically generated

Figure 6 New private variables and constructors

As for critical part with most write operations, I used ReentrantLock to only allow 1 thread to access to it in any given time.

Text

Description automatically generated

Figure 7 Critical part is protected with a lock

### 3.3.2 Parallel stream

### 3.3.3 Executor Service

# 4. Outcomes

d. timing and profiling results, speedup graphs

e. same results (number of promoters identified)?

## 4.1 ExplicitThreading

## 4.2 Parallel stream

## 4.3 Executor Service

# 5. Difficulties

## 5.1 Explicit Threading

The app can save some computation power if ParseReferenceGene is moved down to my new class ExplicitThreading. Because then, the referenceFile can now be processed once in ExplicitThreading and used multiple times by adding the variable as a parameter in every thread call instead of compute it every time a new Thread is call.

However, ParseReferenceGene depends on ‘consensus’ and functions, which makes the program much more complex. While attempting to do this, I got a data race error and did not manage to fix it.

Exception in thread "0" java.lang.NullPointerException: Cannot invoke "qut.Sigma70Consensus.addMatch(edu.au.jacobi.pattern.Match)" because the return value of "java.util.HashMap.get(Object)" is null

at qut.Sequential\_for\_explicit\_threading.run(ExplicitThreading.java:93)

So, this further development is stopped.

## 5.2 Parallel Stream

Have to turn *sigma70\_pattern* to a List of ThreadLocal variable due to the use of parallelStream and executorService. <https://www.baeldung.com/java-threadlocal>

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

# 6. Reflection

# 7. References

<https://www.baeldung.com/java-executor-service-tutorial>

<https://www.baeldung.com/java-when-to-use-parallel-stream>

# 8. Appendixes

parallelStream 111 50s – 76s

A picture containing chart

Description automatically generated

Diagram, timeline

Description automatically generated

101 57s 78s 79s

Diagram

Description automatically generated

Chart

Description automatically generated with medium confidence

011 72 less overhead?

Diagram

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

001 77

Diagram

Description automatically generated

Chart

Description automatically generated

110 83

Thread waiting for other thread to finish

Timeline

Description automatically generated

A picture containing chart

Description automatically generated

100 108

Too many thread waiting

Timeline

Description automatically generated with medium confidence

Timeline

Description automatically generated

010 101s

Only create 5 threads

Timeline

Description automatically generated A picture containing timeline

Description automatically generated

Sheep1 62 77 78 85 50 60

Sheep2 52

Sheep2 PS 12 threads average 63s

Me 8 threads average 65s

Me 12 threads average 70s

Me 8 threads without specify average 84s