You are logged in as Wei Chen (Logout)

CSCI 532: HIGH PERFORMANCE

Home CS532 19 September - 25 September MPI Assignment 2 - DNA Sequence Alignment

Homework 2 is to create a parallel sequence alignment program using MPI.

Your program should be able to take multiple seed (read) input files, and one reference genome files (a chromosome) and print out a single file containing the indices and counts of all the matched reads in those chromosomes. For a simple example, if the read file containeds:

ACGT

AGGT

ATTTT

ATTTT

and the reference genome files were:

>chr1

ATTTTTACGGTAACGTAGGT

the output file would contain:

chr1, 0, 2

chr1, 12, 1

chr1, 16, 1

where the first value on each line is the chromosome, the second is the index where there was a match and the third is the number of reads matching the sequence at that index.

Your program needs to:

1. (20%) Divide the reads across each process. Each process should get a read file or a portion of a read file to work on. The process should take each read and put it into an std::unordered_map<string,int>, where the string is the sequence, and the int is the number of times it occurs in the file. You can ignore any reads with Ns. You can ignore some lines, the data file looks like this:

@ERR192339.1 HWUSI-EAS493_0001:2:1:995:9863/1
NGAGCCGTCACAGCCTGCCGTGGGAAACCTCNCCCCNGNN
+
#)+*)-)),'3-335AAAA887207A55A###########
@ERR192339.2 HWUSI-EAS493_0001:2:1:995:17601/1
NCAGAATTTGCATCATGAACGATGAGCTGATCGTGANGNN
+
#)))'())(+A7AAA0)00.8-8-8AA-AA##########
@ERR192339.3 HWUSI-EAS493_0001:2:1:995:5061/1
NCACAATCTGCTTCCCAGCACTGACAGCCAAGTCACNTNC

...

The lines in bold are the reads, you can ignore the other lines.

NAVIGATION

Home
My home
Site pages
My profile
Current course
CS532
Participants
General
22 August - 28 August
29 August - 4 September
5 September - 11 September
12 September - 18 September
19 September - 25 September
MPI Assignment 2 - DNA Sequence Alignment
26 September - 2 October
3 October - 9 October
10 October - 16 October
17 October - 23 October
24 October - 30 October
31 October - 6 November
7 November - 13 November
14 November - 20 November
21 November - 27 November
28 November - 4 December
5 December - 11 December
12 December - 18 December
19 December - 25 December
My courses

SETTINGS

Course administration
My profile settings

2. (20%) Each process should read the chromosome file after reading the seed file. The easiest way to do this would be to have a std::string and just append each line of the file to it. I would recommend creating some smaller sized test files to start with. If you use the head -n 100 <file> command it will print the first 100 lines of the file to the screen, this is an easy way to get a smaller version of a file.

3. (20%) After the reads have been read into the unsorted_map, and the chromosome file has been read, the process should start at index 0 and check to see if the first 40 characters are in the unsorted_map, if they are it needs to asynchronously send a message to the master process (which will handle writing the sequences to the output file) -- use MPI_Isend for this; it should do this for every index in the chromosome. You can ignore any sequences in the reference genome with Ns.

4. (20%) The master process should repeatedly receive the incoming messages from the other processes, putting the results into another unordered_map, adding to the count if different processes have reads at the same index.

5. (20%) When all the other processes have completed, they should send a finish message to the master process (you can use a different tag for this), and when the master process receives a finish message from all the other processes then it should write the contents of its hashtable to the output file.

Submit your completed homework as a zip file called '<your_last_name>_hw2.zip' to moodle.

BONUS (10%): Have your program handle up to X Ns in each read, and up to Y Ns in each reference genome subsequence. These should be command line parameters.

BONUS (10%): Have your program handle multiple chromosome files as well.

GETTING THE DATA FILES:

You can download these files using wget from the command line. This way you can easily download them to remote accounts, instead of downloading them to your computer and then uploading them to the remote account. You can also use the -c command to resume a partially downloaded file, if the download quit for any reason, for example:

```
wget -c http://hgdownload.cse.ucsc.edu/goldenPath/mm10/chromosomes/chr1.fa.gz
```

will download chr1.fa.gz to the current directory, and continue the download if it was previously broken.

You can get the seed files (the reads) from:

www.ebi.ac.uk/ena

enter in ERP001953

To download them from the command line (from the above website):

```
wget -c ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR192/ERR192339/ERR192339.fastq.gz

wget -c ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR192/ERR192340/ERR192340.fastq.gz

...
```

You can download the mouse reference genome here:

http://hgdownload.cse.ucsc.edu/goldenPath/mm10/chromosomes/

you want to download the files chr*.fa.gz, which are gripped files of each mouse chromosome in FASTA format, e.g.:

```
  wget -c
http://hgdownload.cse.ucsc.edu/goldenPath/mm10/chromosomes/chr1.fa.gz

  wget -c
http://hgdownload.cse.ucsc.edu/goldenPath/mm10/chromosomes/chr2.fa.gz

  ...
```

Available from:	Friday, 12 February 2016, 6:05 PM
Due date:	Thursday, 12 October 2017, 11:55 PM

Upload a file

You are logged in as Wei Chen (Logout)

CS532