Homework #1

Due February 14th, 11:59pm

Each homework submission must include:

- An archive (.zip or .gz) file of the source code containing:
  - The makefile used to compile the code on Monsoon **(5pts)**
  - All .cpp and .h files **(5pts)**
- A full write-up (.pdf of .doc) file containing answers to homework's questions **(5pts)**, including the exact command line needed to execute every subproblem of the homework

The source code must follow the following guidelines:

- No external libraries that implement data structures discussed in class are allowed, unless specifically stated as part of the problem definition. Standard input/output and utilities libraries (e.g. math.h) are ok.
- All external data sources (e.g. input data) must be passed in as a command line argument (no hardcoded paths within the source code **(5pts)**.
- Solutions to sub-problems must be executable separately from each other. For example, via a special flag passed as command line argument **(5pts)**

For this homework, you will need to use the most recent human genome assembly located on Monsoon: /common/contrib/classroom/inf503/genomes/human.txt

- This file contains multiple scaffolds that comprise the human genome
- The genome is in FASTA format (see insert)
  - The headers are unique and always begin with the ">" character. These can be discarded for this homework.
  - Each line of genome file is exactly 80 characters long (plus carriage return character)
  - The genomic sequences consist of the following alphabet {A, C, G, T, N}

```
>568815281-9606
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTCACCCTTCTAACTGGACTCTGACCCTGATTGTTGAGGGCTGCAAAGAGGAA
GAATTTTATTTACCGTCGCTGTGGCCCCGAGTTGTCCCAAAGCGAGGTAATGCCCGCAAGGTCTGTGCTGATCAGGACGC
AGCTCTGCCTTCGGGGTGCCCCTGGACTGCCCGCCCGCCCGGGTCTGTGCTGAGGAGAACGCTGCTCCGCCTCCGCGGTA
CTCCGGACATATGTGCAGAGAAGAACGCAGCTGCGCCCTCGCCATGCTCTGCGAGTCTCTGCTGATGAGAACACAGCTTC
ACTTTCGCAAAGGCGCAGCGCCGGCGCAGGCGCGGAGGGGCGCGCAGCGCCGGCGCAGGCGCGGAGGGGCGCGCCCGAAC
CCGAACCCTAATGCCGTCATAAGAGCCCTAGGGAGACCTTAGGGAACAAGCATTAAACTGACACTCGATTCTGTAGCCGG
CTCTGCCAAGAGACATGGCGTTGCGGTGATATGAGGGCAGGGGTCATGGAAGAAAGCCTTCTGGTTTTAGACCCACAGGA
AGATCTGTGACGCGCTCTTGGGTAGAGCACACGTTGCTGGGCGTGCGCTTGAAAAGAGCCTAAGAAGAGGGGGCGTCTGG
AAGGAACCGCAACGCCAAGGGAGGGTGTCCAGCCTTCCCGCTTCAACACCTGGACACATTCTGGAAAGTTTCCTAAGAAA
```

Problem #1 (of 2): Monsoon account creation and workshop

- **(25pts)** Navigate to NAU's High Performance Computing Cluster (Monsoon) account creation page at https://in.nau.edu/hpc/obtaining-an-account/
- Complete the Self-Paced Workshop
- Obtain and submit the validation codes to self-validate your account
- Take a screenshot of the successful 'confirm user' command (see example below) and submit it as part of your writeup to complete problem #1 of the assignment.

```
$ module load workshop
$ confirm_user
username: abc123
exercise 1 code: 104b0c020063c4b666c376480260f14c
exercise 2 code: b542e82782e8af2663f243d39ec1751c
exercise 4 code: 0659f7fc71a24040c54e8a69c55a38bc
You've successfully confirmed your account!
Press Enter to Exit
```

Problem #2 (of 2): basic text processing

Write code to read, store, and analyze the latest human genome assembly (found at: /common/contrib/classroom/inf503/genomes/human.txt). At minimum, your code must contain (**10pts**):

- A character array to store the entire human genome in a single data structure
- A separate function to read the human genome file
- A function to compute the number of A, C, G, or T characters in the human genome
- Comments describing major code blocks and control structures

A. **(20pts)** Read in and store the human genome. There will be multiple scaffolds (each with a separate header denoted by ">"). Concatenate the entire genome (discard headers) into a single character array data structure. Collect the following statistics (see below) as you are reading the file. Hint: you can keep running totals or store scaffold sizes / names in a separate sets of arrays

   - How many scaffolds were there?
   - What was the longest and shortest scaffold? Provide names of scaffolds and lengths.
   - What was the average scaffold length?

B. **(20pts)** Write a function to assess the content of the human genome – count the total number of a given character (A, C, G, or T) in the whole genome.

   - What is the 'big O' notation of your search (linear / quadratic / cubic / etc)?
   - How long does it take (in seconds) to execute this function? Hint: You will need to use system time within your code to get accurate time estimates.
   - What was the GC content of the human genome (percent of C's and G's in the genome)?