IC 150 P Computation for engineers lab

Lab assignment sheet no: 8, Odd semester, 2016

Reading from a file, Sorting, and searching

Prepared by: Tulika Srivastava

Introduction

Computational approaches are routinely used for solving problems in Life Sciences. Some typical tasks include looking for known simple and complex patterns in given DNA or protein sequences. Simple patterns are easier to make, however complex patterns can be simpler or complex and made using regular expression. Another routine operation is to sort the DNA and protein sequences with respect to their lengths (character counts).

Objective for this lab session

- In this lab session, you will learn to read from a file, perform simple searching of a pattern, and sorting. Download the file ‘‘`seqs.faa`’’ from the moodle page for IC 150 P.

Task:

You are given with 5 protein sequences in the file ‘‘`seqs.faa`’’. ‘‘`.faa`’’ is a standard format in which the DNA or protein sequences are stored for performing computations on them. Each line starts with a ”>” symbol and reports the annotation of the protein or DNA, such as its function, cellular location, organism source, etc. The sequence starts from the next line after the annotation line.

Part 1: Read the sequences from the file one by one. Calculate the length of each protein sequence and store it in an array. Using the function `qsort` from `stdlib.h`, sort the array to determine the lengths of the smallest protein and the largest proteins. Report these values on the standard output. You can use the following code to read from a file:

```c
#include <stdio.h>

#define MAX_STRING_SIZE 1000

int main()
{
    char buffer[MAX_STRING_SIZE];
    FILE *f = fopen("textFile.txt", "r");
    fgets(buffer, MAX_STRING_SIZE, f); /* reads first line, and changes file pointer so that it points to next line */
    printf("Line read: %s\n", buffer);
    fclose(f);
    return 0;
}
```

readFirstLineFromFile.c

Note: Every instance of the `fopen` command such the one on line number 8, must be matched by a matching instance of the `fclose` command (as on line number 11). Once you have created a file pointer using `fopen`, you do not need to call `fopen` again for reading from the same file. To read more lines from the opened file, you just have to use `fgets` a few number of times. Experiment with `fgets` to see how to read lines with specific line numbers.

Part 2: Use the following simple pattern:

```
VTEVGIFTPKAVGR
```

to search in each of the 5 protein sequences given in the file ‘‘`seqs.faa.`’’ Print on STDOUT the annotation of each sequence and the start and end location of the pattern where it occurs in this sequence. Repeat this for all the five sequences.