CMSC 423 Fall 2012: Project Specification

Introduction

The project will consist of four components due throughout the semester (see below for timeline). Basic rules:

- You are allowed to work in teams of at most 2 people.
- The teams can change throughout the semester (i.e. you can work on part 1 with one of your colleagues and on part 2 with another one if you wish). Please clearly indicate on your submission who the members of the team are (both will get the same grade, irrespective of contribution).
- You can use any programming language you wish.
- Your software must compile and run on the Glue machines so make sure that you test it before submitting it. **You can get at most 50% of the grade if we have difficulties compiling or running your code.**
- The projects must be submitted using the "submit" command on the Glue system. Note: this is different from the "submit server". You must be logged onto glue.umd.edu or linux.grace.umd.edu in order to run the "submit" command.
- Your code must be accompanied by a README file that explains the steps necessary to compile and run your project.
- 10% of the grade for each component of the project will be awarded for "best programming practices" - make sure your code is neat, well organized and thoroughly commented.

Deliverables/Timeline

- FASTA parser.
  Due: 9/20/12 Weight: 10%
- Global alignment of two DNA sequences.
  Due: 10/25/12 Weight: 20%
- Local alignment with affine gap penalties.
  Due: 11/08/12 Weight: 30%
- Overlapper for assembly & incorporation into Minimus assembler.
  Due: 12/4/12 Weight: 40%

Part 1 - FASTA parser
Due Thursday, September 20, 2012
Overall weight: 10% of total project grade

The first part of the project requires you to write code that can parse a FASTA file. For this part, you are not allowed to use any of the Bio* libraries available for your programming language of choice.

Specification: Your program should read in a FASTA file (sample is available on the Glue system in /class/fall2012/cmsc/423/0101/public/test.fasta) and output a list of sequence identifiers for all sequences that satisfy one of the following:

1. are less than 100 bp in length
2. contain at least one character that is not A,C,T, or G

Details:

- For clarifications on the FASTA format see the Wikipedia entry: http://en.wikipedia.org/wiki/FASTA_format
- In addition, you can assume that a sequence identifier follows right after the ">" sign. If a sequence doesn't follow this rule you can exit with an error.
- You can also assume that the identifier ends with the first "space" character (space, tab, or end-of-line)
- **Interface:** Your program must accept the input fasta file either through the standard input, or as the only command-line parameter. The output should be provided on standard output.
- Any questions about this assignment should be sent to both myself and the TA.