Lab 2-1: DNA Analysis: GC Percent

Due date: April 3/April 5.

About the Lab

This is your first software development lab in the course. One part of the lab is due by the end of the lab period, while the full set of lab deliverables is due by **April 5**. Please note, that **Lab 2-2** will start during the April 5 lab period, so your deliverables for **Lab 2-1** must be submitted by the beginning of the class.

The lab is a **joint group assignment**. You will perform this lab in the **CSC 448/CHEM 441** teams assembled during the first week of classes.

Please note that the **CSC 448** and **CHEM 441** deliverables for this lab are different. Your goal is to develop software that allows your team partners from **CHEM 441** answer their questions.

This lab illustrates the process of working on **multidisciplinary teams** in miniature. Because you only have one lab period to work on this lab, the timing is important. Here is a rough breakdown of the lab period time for you.

Time	CSC 448	CHEM 441
10 mins	Studying assignment	Studying assignment
10 mins	Discuss assignment/map out solution	
30 mins	Software development	Data collection
20 mins	Assembly, testing, modification, and use of software	
10 mins	Preparing and submitting deliverables	

Lab Assignment

Recall from the lectures that DNA is the genetic material of life that encodes instructions for making building blocks of living cells proteins. DNA is a linear polymer that contains information in the order/sequence of bases. Four bases found in DNA are G (guanine), C (cytosine), T (thymine), and A (adenine). In DNA molecules, G is always paired with C on the opposite strand, while T is always paired with A.

GC-content. GC content also known as GC percentage is defined as the percentage of DNA bases that contain either G (guanine) or C (cytosine). The frequency of bases in a DNA polymer varies among different organisms. GC content of an entire genome (all DNA of one organism or species) can be used to distinguish two closely related species or to find a foreign gene in a genome (a gene that jumped species; for example, viruses can transfer genes from one organism/species to another). In addition, GC content can vary between different functional regions within one genome. When analyzing a new genome, finding regions of high or low GC percentage can point to particularly interesting regions (e.g. protein coding genes, origin of replication). Most molecular biologists use GC content of small regions of DNA to choose primers (molecular tools) for DNA amplification by PCR.

Programming assignment. The core of your assignment is development of code that computes the GC content of a given DNA sequence. This code, however, comes with a number of non-functional requirements:

1. The **CHEM 441** students on your team received a specific research question that they need to address by comparing GC content of various DNA sequences. They **will rely on the software you write** to obtain the necessary information. Therefore, your software must address their specific needs.

2. You must reach agreement with **CHEM 441** members of your team on

   - the inputs to your software;
   - the output your software produces;

3. Upon delivery of the software to **CHEM 441** members of your team, you must train them how to run and use the software. Note, that for the in-class part of the assignment, there is **no expectation** of any graphical UI: everything can be command line, but **CHEM 441** student must learn how to use the tools you give them.

The **CHEM 441** members of your team are responsible for providing you with the input data.

April 3 deliverables. Submit the GC content computation program you developed for the **CHEM 441** memebers of your team. The full list of files to submit:

1. All files for your program.
2. `README` file that contains the name of your team, names of all **CSC 448** and **CHEM 441** team members, and instructions on compiling and running your program.
3. Data files on which the program is to be run.

April 5 deliverables. We are making code for a generic simple Java front end available to all teams. In completing this task you can either use this code, or you can build your own GUI. Your task is to create a simple UI-driven tool that allows **CHEM 441** students to compute GC content of their data. The tool should allow them to load necessary data files. The tool can actually display the computed results, or it can dump the results into a human readable file (a .txt or a .csv file, for example), and simply inform the users that the work is complete.

Make sure that you share the tool with your **CHEM 441** team members prior to submitting it and make sure that they are comfortable using it. Submit all the necessary source code for the tool. You need to generate an executable that works on the computers your **CHEM 441** teammates are planning to use for their study and to deliver that executable to them.

You need to submit all files used to create the software. In addition, put together another `README` file (name it `README-April5` which puts describes how to compile and run your code.

Submission Instructions

These instructions are for your graded deliverables for **CSC 448**. **CHEM 441** students have their own set of deliverables: they rely on being able to run your software to produce them.

Use handin to submit all your files. Submit your **April 3 deliverables** one file at a time. Submit your **April 5 deliverables** except for the `README-April5` (which you submit separately) file as a single `.zip` or `.tar.gz` archive.

Use the following command to submit:

```
$handin dekhtyar 448-lab2-1 <files>
```