Lab 4: Sequence Alignment

Due date: May 22/24.

About the Lab

This is your last joint **programming** lab with **CHEM 441**. The **CSC 448** teams remain the same as in prior labs. The pairing with the **CHEM 441** students remains the same.

This lab completes the construction of the software suite for comparative analysis and annotation support of fruit fly genome. There is only one deliverable for this lab: the software to be used by the **CHEM 441** students.

Time	CSC 448	CHEM 441
May 15 lab	Discuss assignment/map out solution	
May 17 lab	Work on implementation	
May 17 – May 22	Software development	Data preparation
May 22 lab	Delivery of working prototype	Prototype use
May 24 lab	Software delivery	Software use

Lab Assignment

The software you developed for **Lab 3-1** allows **CHEM 441** students to combine into a single contguous string a group of consecutive DNA fragments (contigs). In that lab, each contig came with a GFF file specifying the locations of the coding regions. One of the goals of **Lab 3-1** was to build the GFF file for a the merged DNA fragment by recomputing the coordinates of each coding region in the newly constructed DNA string.

For each coding region belonging to the intersection of two consecutive contigs, there should have been (and the were) two entries in the respective GFF files. Upon converting the entries to the new coordinates, your **Lab 3-1** code was supposed to check if the two entries coincided and report any conflicting coding region descriptions.

Lab 4 builds on these reports. The goal of the lab for the **CHEM 441** students is to determine which of the conflicting annotations should be kept, and which should be removed from the final GFF file.

The more formal definition of the problem that needs to be solves is as follows. Consider a DNA string S representing a fragment of a genome of a known organism (the data you will be using is for *Drosophila Erecta*, a species of the fruit fly). Consider an *exon* e of some gene g found inside S, for which two conflicting ranges, $[x_1, y_1]$ and $[x_2, y_2]$ are given.

The goal is to determine which of the two ranges best represents the exon in question.

The process for doing so uses a reference genome a known established DNA sequence for the exon in question, from a related species. For **CHEM 441** students, the reference species is *Drosophila Melanogaster*.

Given the reference DNA string for an exon, and the two *candidate exons*, to determine which candidate exon annotation is kept and which is removed, the two candidates exons needs to be compared to the reference string. The exon that is more similar is kept, the other one is removed.

The final decision-making on which exons to purge and which to keep belongs to your **CHEM 441** lab partners. However, in doing so, they will rely on the software that you will build for them, that compares two DNA strings and establishes their similarity.

The **specific information** on how to do so **shall be delivered to you** by your **CHEM 441** partners during the May 15 lab period in the form of written requirements.

May 15 lab. During this lab period, discuss the requirements with your **CHEM 441** lab partners. Please make sure you understand all aspects of the assignment, and you understand, which specific algorithms you need to implement in order to compare two DNA sequences for similarity, and what output your **CHEM 441** partners are expecting.

Please, contact me during the lab if the requirements provided

by the CHEM 441 students are insufficient, incomplete, or unclear. This lab has a very short time span, so it is our goal to ensure that by the end of the May 15 lab **every CS team** understands what it needs to implement.

Deliverables

There is only one deliverable for this assignment: the software that compares two DNA strings for similarity. The software will wind up being an implementation of one of the algorithms discussed in class, wrapped with code that parses the input, and prepares and delivers the output.

As with every piece of software you are writing for your **CHEM 441** partners, you need to negotiate the user interface in advance. The specifics (command line, text-based menus, GUI, etc...) are left to each team. The only requirement is that **CHEM 441** students are comfortable using the tool.

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.

On **May 22** you must deliver a prototype of your software tool to your **CHEM 441** partners. The prototype must perform correct computations and produce correct results, but it need not possess the final polished UI and/or possess the finalized I/O capabilities. The prtotype shall be usable (if not by your **CHEM 441** partners independently, then, at least, by you) and its results shall be sufficient for **CHEM 441** students to investigate exon similarity between the species they are studying and the reference species.

No formal prototype submission other than delivery to **CHEM 441** students is requried.

On **May 24** the final version of the software needs to be delivered and submitted. Submit using handin. Include a `README` file with explanations on how to run your software (and the names of all team members).

Use the following command to submit:

```
$handin dekhtyar 448-lab4 <files>
```