Lab 6: Contig Assembly and Quality Control

Due date: Monday, June 3.

About the Lab

This is your last joint **programming** lab with **BIO 441**. This lab completes the construction of the software suite for comparative analysis and annotation support of fruit fly genome.

Time	CSC 448	CHEM 441
May 23 lab	Discuss assignment/map out solution	
May 23 – May 29	Software development	Data preparation
May 29 lab	Delivery of working prototype	Prototype use
May 30 – June 3	Testing/Refinement	Testing/Use
May 24 lab	Software delivery	Software use

Lab Assignment

For comparative genomics analysis biologists need to look at fairly large regions of DNA sequence: 200Kb 2Mb. Our data comes in 30-60Kb segments of DNA sequence. The segments are named contigs or fosmids and are numbered sequentially: *D. erecta* contigs 1-38 are from the Dot/4th chromosome, fosmids 1-42 are from 3L control region. The terms contig and fosmid have been borrowed from the genome sequencing terminology (you can look up real meaning of the terms on the web); in our case, they just designate specific DNA segments. The contigs partially overap, so the end sequence of one contig will match the beginning of another, and overlapping regions contain the same genes.

In order to analyze larger regions of genome, **BIO 441** students need to combine multiple annotated contigs into larger segments (supercontigs) and produce a new FASTA sequence that contains the combined DNA sequence accompanied by a new GFF file that contains new coordinates.

This task requires development of the software that completes two actions:

1. **Sequence combination.** Combines provided FASTA files for the contigs into a single supercontig file.
2. **Annotation combination.** Combines provided GFF annotation files for individual contigs and produces a single GFF annotation file representing all gene annotations on the supercontigs constructed.

Sequence combination

Two consecutive contigs C_i and C_{i+1} overlap by a number of nucleotides. Essentially, the suffix of C_i is the prefix of C_{i+1}:

$$C_i = a_1 \dots a_N b_1 \dots b_K$$

$$C_{i+1} = b_1 \dots b_K c_1 \dots c_M$$

Merging together two contigs C_i and C_{i+1} means finding the largest overlap between C_i and C_{i+1} and "gluing" the two strings together to produce a string $C' = a_1 \dots a_N b_1 \dots b_K c_1 \dots c_M$. This process needs to be repeated for all consecutive contigs. The requirements provided to you by your **BIO 441** partners should outline how to handle the cases of missing contigs (there are some that are missing in the data).

Please note, that the contigs may contain the `"N"` characters that can represent multiple nucleotides. Your requirements documents should contain specific instructions on how these are to be handled.

This part of the assignment primarily uses the techniques from the last couple of labs. This is yet another chance for you to try your hand at implementing Ukkonen's algorithm.

Annotation combination

Combining annotations from multiple GFF files has two key aspects that need to be considered.

Coordinate recomputation. Each original GFF file reports annotations in the coordinates within the frame of the FASTA sequence being annotated. When contigs are merged, all coordinates for contigs 2 and above must shift. It is your responsibility to compute the appropriate shifts and transfer the gene annotationg into the GFF annotation file for the supercontig with appropriate coordinates.

Quality control. Different contigs were annotated by different people. Some of the annotations may appear in the overallping regions of two neighboring contigs. You must determine if annotations in the overlapping regions are consistent. If they are not (two annotations refer to the same gene/exon, but resolve to different coordinates within the supercontig), your software must determine which annotation is better.

The requirements provided to you by your **BIO 441** partners should contain detailed instructions for how this dermination must be made. On the technical side, it boils down to finding a *reference sequence* from an already annotated genome of a related species (*D. melanogaster* in the case of the fruit fly genomes, usually, and determining which of the annotations of *D. erecta* is *more similar* to it. This is done using some form of sequence alignment.

The requirements provided to you by your **BIO 441** partners will contain the explanation of the steps to take when genome annotation conflicts are detected.

Submission Instructions

As usual, two sets of deliverables are needed: one for `handin` and one - for `Piazza`. The core deliverables are the code and two requirements documents: the original document provided to you by your BIO 448 partners and the final version.

handin deliverables. The `handin` deliverables are: *source code*, *compilation/running instructions (README)*, *user documentation* and *both requirements documents*. Submit them using the following command:

```
$handin dekhtyar 448-lab6 <files>
```

Deadlines. There is only one deadline: submit everything by the end of the calendar day on Monday, June 3. Please note, that your **BIO 441** partners may want to work with preliminary versions of the software before the submission deadline.